Gamma Ray Point Source Data Description

March 17, 2021

Introduction

The data of this repository contains simulations of the γ-ray sky, which are oriented to reproduce or compare to the physical observables obtained by the Fermi-LAT telescope over 9.5 years of data acquisition. These simulations are represented as images (matrices), where each pixel of the image (matrix cell) represents a position in the sky and energy of the photons. The content of each pixel is the photon number (as a floating point number in the Asimov limit) associated to the emission from different γ-ray sources, such as the interstellar medium (IEM) and different types of point sources (PSs).

In order to reduce the size and complexity of full sky representations, each image is further divided in smaller patches. For instance, concerning point source localization tasks, we consider medium size patches that cover $10^\circ \times 10^\circ$ (64×64 pixels) regions of the sky. Each of these patches include the contribution from three sources. On the one hand we have the diffuse contribution from the simulations of the IEM plus the extragalactic isotropic γ-ray background (IGRB). On the other hand we consider contributions from the two main classes of point sources detected by Fermi-LAT, which are given by Active Galactic Nucleai (AGN) and Pulsars (PSR). Furthermore, each patch is also divided in 5 energy bins (0.5-1, 1-2, 2-7, 7-20, $>$20 GeV). Thus, in practice, the patches dedicated to localization exercises are given by numpy arrays with shape $(3, 64, 64, 5)$.

Concerning the point source classification task, we consider small size patches that approximately cover $1^\circ \times 1^\circ$ (7×7 pixels) regions of the sky. In this case we do not distinguish between different contributions to the photon counting, rather we keep the division in terms of energy bins, such that the point source patches are given by numpy arrays with shape $(7, 7, 5)$.

The provided data sets are meant to be used to compare various algorithms, but are not meant to be applied to the real data. We are preparing a new data set that we will release in the near future, that will be optimised for real data applications.

Data content details

The files associated to the patches for localization and classification tasks are released in three different data sets, which are described as follows:

- `patches_localization_training`: this folder contains around 120.000 medium size patches covering almost 300 instances of the full sky. For each image patch there is a corresponding mask patch, which is a numpy array file with shape $(64, 64, 2)$, that contains binary information about the presence of point sources in the corresponding patch. Basically, for each point source we draw a solid disk of radius 2.5 pixels that distinguish IEM regions (pixel value equal to 0) and point source ones (pixel value equal to 1).

- `patches_classification_training`: this folder contains approximately 62.000 small size point source patches, covering around 25 instances of the full sky. Each point source patch is obtained from boxes of dimension $7 \times 7 \times 5$ centered around the positions predicted by the localization algorithm called UNEK (see associated paper). The type of the source, AGN or PSR, is obtained from the true information. The idea behind the construction of this dataset is to define point sources which are more realistic, such that we can directly classify point sources generated from the localization algorithm. For instance, this set of point sources includes boxes centered at positions with a slight off-set in comparison to the true position of the sources. Also, we include FAKE sources which are generated from predictions of the UNEK that do not match a true source, so these FAKE sources are mostly made from background with a small contamination from near true sources.

- `patches_test`: this folder contains 5 sub-folders of medium size patches that cover different realizations of one single instance of the full sky. The sets F0-B1, F0-B2 and F0-B3 cover three different versions of the background, while the sets F0-B1, F1-B1 and F2-B1 cover three different realizations of the point source content distribution (for details see the associated paper). These folders do not contain the associated masks or CSV files, thus in order to evaluate the predictions of future algorithms considering this data, the interested user must contact to the responsible of the dataset in order to access to the true information regarding these patches.

Along with the previous folders, that only contain numpy format files, we also release CSV files containing the true information of the training samples. Basically, each row of the CSV files (localization and classification) includes the information associated to each point source contained in the respective patch (several sources for localization patches or a single one for classification patches). For completness, let us summarize the content (columns) of each of these files, starting with the localization CSV files (training and validation):

- $filename$: file name of the corresponding medium size patch, which is a numpy format file with shape (3,64,64,5)
- x_{min}, x_{max}, y_{min}, y_{max}: pixel coordinates of the 5×5 box around point source positions. This is the square that is circumscribed around the disk that define each point source position
- $class$: true class of the point source, we use (0,1) for (AGN, PSR)
- lon_c, lat_c: galactic coordinates of the center of the patch, $lon \in [-180, 180]$ and $lat \in [-90, 90]$
- $flux$ (S_1): photon flux computed from 1 GeV to 100 GeV in units of [photons $\mathrm{cm}^{-2}\mathrm{s}^{-1}$]. This flux corresponds to the nearest true point source, for which we know the information from the simulation parameters
- lon_p, lat_p: galactic coordinates of the true point source
- $catalog$: full sky instance number, $catalog \in [0, 1000]$

The columns of the classification CSV files (training and validation) are as follows:

- $filename$: file name of the small size patch associated to a single point source, which is a numpy format file with shape $(7, 7, 5)$
- lon_{ps}, lat_{ps}: predicted position in galactic coordinates of the corresponding point source (UNEK output)
- $catalog$: full sky instance number
- $test_{\mathrm{Flux,1000}}$ (S_1): photon flux computed from 1 GeV to 100 GeV in units of [photons $\mathrm{cm}^{-2}\mathrm{s}^{-1}$]. This flux corresponds to the nearest true point source, for which we know the information from the simulation parameters
- $class$: true class of the nearest true source. We use (0,1,2) = (AGN, PSR, FAKE), where the class FAKE is used when the nearest true source is farther than 0.3 degrees
- $test_{lon}$, $test_{lat}$: position in galactic coordinates of the nearest true point source
- $test_x$, $test_y$: pixel coordinates position of the nearest true point source
- $distance_{dg}$, $distance_{px}$: distance in degrees and pixels between the predicted and true point source positions
- $test_{\mathrm{Flux,10000}}(S_{10})$: photon flux computed from 10 GeV to 1 TeV in units of [photons $\mathrm{cm}^{-2}\mathrm{s}^{-1}$]. This flux corresponds to the nearest true point source, for which we know the information from the simulation parameters

Secret test protocol

As we have mentioned in the previous sections, in this repository we release 5 data sets that do not include the corresponding true information. The purpose of these tests is to stablish a centralized approach to measure the performance of new localization and classification algorithms based on the training data released in the current repository. To know your performance, please send us a CSV file named as the corresponding test set containing the following columns: $filename, x_{ps}, y_{ps}, class$, where $filename$ is the file name of the analyzed medium size patch, x_{ps} and y_{ps} the cartesian coordinates of the predicted point source and $class$ is the type of source, 0=AGN, 1=PSR or 2=FAKE.