FAQ (Frequently Asked Questions):

Author: Ke Yan, ke.yan@nih.gov

Imaging Biomarkers and Computer-Aided Diagnosis Laboratory

National Institutes of Health Clinical Center

Version 5, 05/13/2019

1. How to download the 56 zip files in batches?

Please run the provided script: batch_download_zips.py

2. I cannot open the downloaded zip files.

The zip files were created using the zip command on Linux. If there are errors, they may be caused by the incompleteness of the downloaded files. Please compare their MD5 checksums with those in Images_png/MD5_checksums.txt, or try to download the files again from different computers and places.

3. The images in folder "Images_png" have very low contrast.

Note that the images in "Images_png" are 16-bit images. To view them, you may:

- Convert the png files to 3D nifti files using our provided python script, then view nifti in software such as 3D slicer and ITK-snap;
- Write a script to process the images:
    1) Convert the image to int32 format, then subtract 32768 from the pixel intensities to obtain the original Hounsfield unit (HU) values (generally about -1000 ~ 1000, https://en.wikipedia.org/wiki/Hounsfield_scale);
    2) Do intensity windowing (https://radiopaedia.org/articles/windowing-ct) on the HU values, i.e., convert the intensities in a certain range ("window") to 0-255 for viewing. To view different structures (lung, soft tissue, bone etc.), we need different windows. The column "DICOM_windows" in DL_info.csv provides the default window for each image. For example, if the min and max values of a window is A and B, then the windowed intensity I should be
    $$I = \min(255, \max(0, (HU-A)/(B-A)*255);$$
    3) Save the windowed image to 8-bit image files. This is how the files in Key_slices.zip were generated.

4. What are the naming conventions of the images?

We named each slice with the format "{patient index}_{study index}_{series index}_{slice index}.png" , with the last underscore being / or \ to indicate sub-folders, depending on the operating system. Note that one patient often underwent multiple CT scans (studies) for different purposes or follow-up. Each study contains multiple volumes (series) that are scanned at the same time point but differ in image filters, contrast phases, etc. Every series is a 3D volume composed of tens to hundreds of axial image slices.

5. Why are only part of the slices provided in each volume?

The DeepLesion dataset was collected based on radiologists' annotations called "bookmarks". We provide key slices that contain the bookmarks as well as at least 60 mm contexts (extra slices above and below the key slice) to facilitate usage of 3D information. Currently we have no plan to release the whole volumes because the data size will be too big.

6. Are the radiological reports included in the publicly accessible DeepLesion dataset?

No. We currently do not have a plan to release the radiological reports. However, we have provided the semantic labels (tags) for the lesions mined from radiological reports. The labels describe the lesions' body part, type, and attributes. Please find them at https://github.com/rsummers11/CADLab/tree/master/LesaNet

7. Are there any restrictions in using this dataset?

The usage of the data set is unrestricted. However, it is recommended to cite our JMI 2018 paper and provide the link to our original download site in your paper.

8. Are the images available in DICOM format?

We decided to provide the losslessly compressed 16-bit png images currently, which are anonymized, more compact, and bear no information loss. Please subtract 32768 from the 16-bit pixel intensities to obtain the original Hounsfield unit (HU) values.

9. What are the definition of "lesions" in the dataset?

DeepLesion is built upon RECIST bookmarks in NIH's PACS, which are annotations marked by radiologists during their daily work to measure target image findings; see our paper. We did not further filter these annotations after obtaining them, so they reflect what the radiologists thought were important and measurable findings. Among them, there are commonly studied lesion types such as lung nodules, enlarged lymph nodes, liver tumors, and so on, as well as many less common ones such as bone and soft tissue lesions. Thus, it is a dataset with great diversity.

It should be noted that not all lesions were annotated in the images because radiologists typically mark only representative lesions in each study. There are also a small proportion of the bookmarks that are actually measurements of normal structures, such as lymph nodes of normal size.

10. What do the values in the "Possibly_noisy" column in "DL_info.csv" mean?

We manually checked several things for each lesion, for example, whether the two RECIST diameters intersect, the width-height-ratio of the bounding-box, the mean and standard deviation of the pixels inside the box, and so on. 35 lesion annotations were found to be noisy. A few of them are imaging phantoms. They are marked with 1 in the "Possibly_noisy" column and can be omitted in analysis.

11. What are the definition of the 8 lesion types?

The lesion types are coarsely defined and just for reference. Please find more comprehensive and fine-grained lesion types in the answer to Question 6.

1) Bone
2) Abdomen: lesions in the abdominal cavity that are not in liver or kidney
3) Mediastinum
4) Liver
5) Lung
6) Kidney
7) Soft tissue: miscellaneous lesions in the body wall, muscle, skin, fat, limbs, head, and neck
8) Pelvis

12. Is there any official data split?

The official split randomly generated in patient level is in the 18th column of DL_info.csv, where train=1, validation=2, test=3. It can be used in tasks such as lesion detection, retrieval, etc.