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

Imaging Biomarkers and Computer-Aided Diagnosis Laboratory

National Institutes of Health Clinical Center

Version 5, 09/05/2018

Introduction

The **DeepLesion** dataset contains 32,120 axial computed tomography (CT) slices from 10,594 CT scans (studies) of 4,427 unique patients. There are 1–3 lesions in each image with accompanying bounding boxes and size measurements, adding up to 32,735 lesions altogether. The lesion annotations were mined from NIH's picture archiving and communication system (PACS). Some meta-data are also provided. The contents include:

- Folder "Images_png": png image files. 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. The images are stored in unsigned 16 bit. One should **subtract 32768** from the pixel intensity to obtain the original Hounsfield unit (HU) values.

  We provide not only the key CT slice that contains the lesion annotation, but also its 3D context (30mm extra slices above and below the key slice). Due to the large size of the data and the file size limit of the website, we packed them to 56 smaller zip files for downloading. Please run batch_download_zips.py to download them in batches.
- Key_slices.zip: key slices with overlaid lesion annotations for review purposes.
- Folder "Key_slice_examples": random image examples chosen from Key_slices.zip.
- DL_info.csv: The annotations and meta-data. See Section "Annotations" below.
- DL_save_nifti.py: demo python codes that can convert the provided 2D 16-bit png images to 3D nifti sub-volumes.

Reference

If you find the dataset useful for your research projects, please cite our JMI 2018 paper:

- Ke Yan, Xiaosong Wang, Le Lu, Ronald M. Summers, "DeepLesion: Automated Mining of Large-Scale Lesion Annotations and Universal Lesion Detection with Deep Learning", Journal of Medical Imaging 5(3), 036501 (2018), doi: 10.1117/1.JMI.5.3.036501

The following paper and code are also related with the dataset:

- Ke Yan, Xiaosong Wang, Le Lu, Ling Zhang, Adam Harrison, Mohammadhadi Bagheri,

Ronald M. Summers, "Deep Lesion Graphs in the Wild: Relationship Learning and Organization of Significant Radiology Image Findings in a Diverse Large-scale Lesion Database", IEEE CVPR, pp. 9261-9270 (2018), https://arxiv.org/abs/1711.10535

- Ke Yan, Mohammadhadi Bagheri, Ronald M. Summers, "3D Context Enhanced Region-based Convolutional Neural Network for End-to-End Lesion Detection", MICCAI, 2018, https://arxiv.org/abs/1806.09648
    - https://github.com/rsummers11/CADLab/tree/master/lesion_detector_3DCE
    - This code project includes basic operations you can do with DeepLesion, such as loading annotations and 16-bit png files, intensity windowing, pixel spacing normalization, and training a lesion detector.
- Jinzheng Cai*, Youbao Tang*, Le Lu, Adam P. Harrison, Ke Yan, Jing Xiao, Lin Yang, Ronald M. Summers, "Accurate Weakly-Supervised Deep Lesion Segmentation using Large-Scale Clinical Annotations: Slice-Propagated 3D Mask Generation from 2D RECIST", MICCAI, 2018, https://arxiv.org/abs/1807.01172
- Youbao Tang, Adam P. Harrison, Mohammadhadi Bagheri, Jing Xiao, Ronald M. Summers, "Semi-Automatic RECIST Labeling on CT Scans with Cascaded Convolutional Neural Networks", MICCAI, 2018, https://arxiv.org/abs/1806.09507
- Ke Yan, Le Lu, Ronald Summers, "Unsupervised Body Part Regression via Spatially Self-ordering Convolutional Neural Networks", IEEE ISBI, 2018, https://arxiv.org/abs/1707.03891
    - https://github.com/rsummers11/CADLab/tree/master/body_part_regressor

Annotations

In DL_info.csv, each row is the information of a lesion in DeepLesion. The meaning of the columns are:

1. File name. Please replace the last underscore with / or \ to indicate sub-folders.
2. Patient index starting from 1.
3. Study index for each patient starting from 1. There are 1~26 studies for each patient.
4. Series ID.
5. Slice index of the key slice containing the lesion annotation, starting from 1.
6. 8D vector, the image coordinates (in pixel) of the two RECIST diameters of the lesion. [x_{11}, y_{11}, x_{12}, y_{12}, x_{21}, y_{21}, x_{22}, y_{22}]. The first 4 coordinates are for the long axis. Please see our paper and its supplementary material for further explanation.
7. 4D vector, the bounding-box [x_1, y_1, x_2, y_2] of the lesion (in pixel) estimated from the RECIST

diameters, see our paper.

8. 2D vector, the lengths of the long and short axes. The unit is pixels.
9. The relative body position of the center of the lesion. The z-coordinates were predicted by the self-supervised body part regressor. See our paper for details. The coordinates are approximate and just for reference.
10. The type of the lesion. Types 1~8 correspond to bone, abdomen, mediastinum, liver, lung, kidney, soft tissue, and pelvis, respectively. See our paper for details. The lesion types are coarsely defined and just for reference. Only the lesions in the val and test sets were annotated with others denoted as -1.
11. This field is set to 1 if the annotation of this lesion is possibly noisy according to manual check. We found 35 noisy annotations out of 32,735 till now.
12. Slice range. Context slices neighboring to the key slice were provided in this dataset. For example, in the first lesion, the key slice is 109 and the slice range is 103~115, meaning that slices 103~115 are provided. For most lesions, we provide 30mm extra slices above and below the key slice, unless the long axis of the lesion is larger than this thickness (then we provide more) or the beginning or end of the volume is reached.
13. Spacing (mm per pixel) of the *x*, *y*, and *z* axes. The 3rd value is the slice interval, or the physical distance between two slices.
14. Image size.
15. The windowing (min~max) in Hounsfield unit extracted from the original DICOM file.
16. Patient gender. F for female and M for male.
17. Patient age.
18. Official randomly generated patient-level data split, train=1, validation=2, test=3.

Applications

DeepLesion is a large-scale dataset that contains a variety types of lesions. It can be used for lesion detection, classification, segmentation, retrieval, measurement, growth analysis, relationship mining between different lesions, etc.

Limitations

Since DeepLesion was mined from PACS, it has a few limitations:

- DeepLesion contains only 2D diameter measurements and bounding-boxes of lesions. It has

no lesion segmentation masks, 3D bounding-boxes, or fine-grained lesion types. Therefore, some applications (e.g. lesion segmentation) may need extra manual annotations.

- Not all lesions were annotated in the images. Radiologists typically mark only representative lesions in each study. Therefore, some lesions remain unannotated.
- According to manual examination, although most bookmarks represent abnormal findings or lesions, a small proportion of the bookmarks are actually measurement of normal structures, such as lymph nodes of normal size.

We encourage our fellow researchers / radiologists to share their own annotations on the dataset to benefit the medical imaging community.

Data visualization

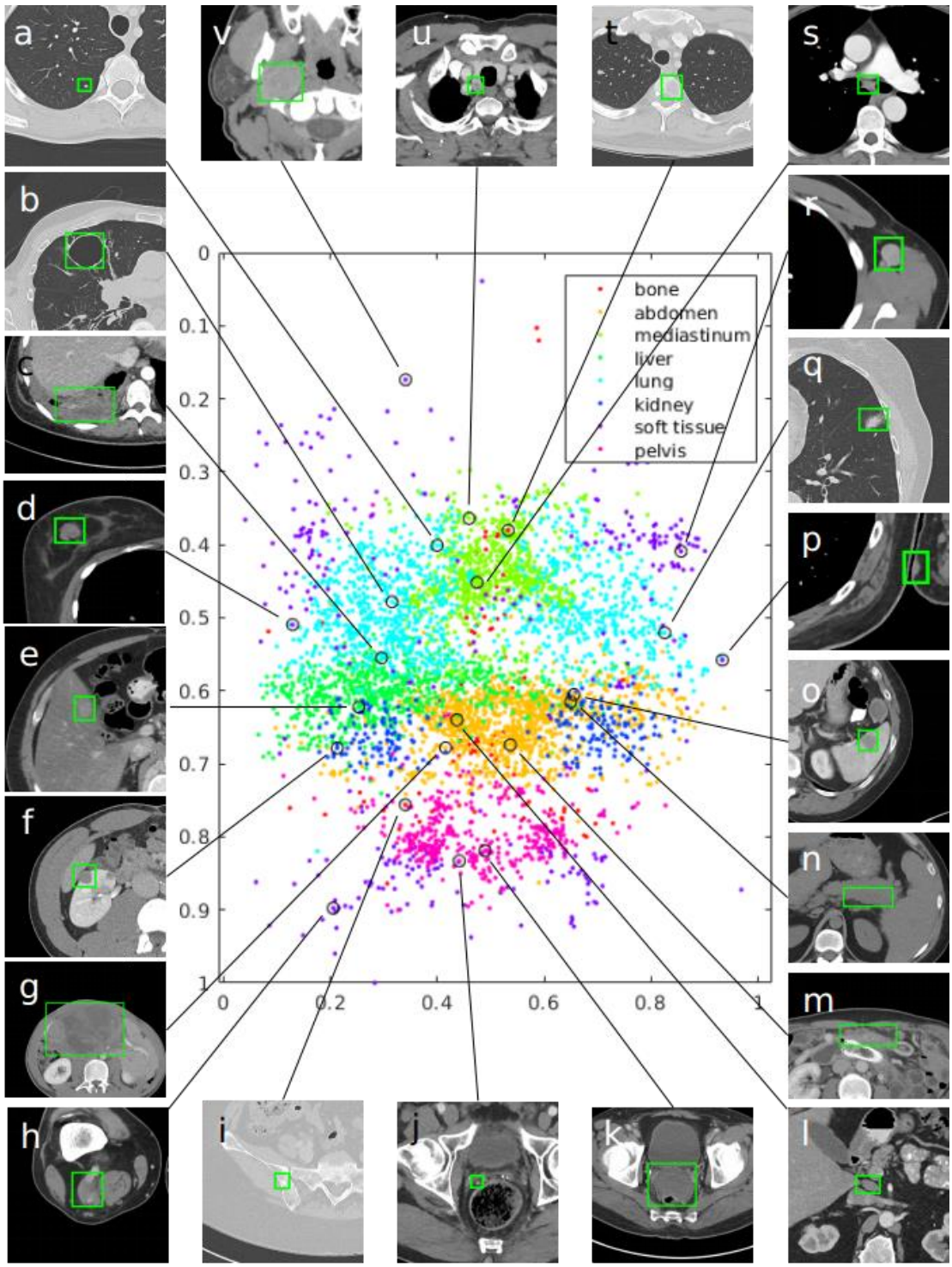

Fig. 1 Visualization of a subset (15%) of the DeepLesion dataset. The *x*- and *y*-axes of the scatter map correspond to the *x*- and *z*-coordinates of the relative body location of each lesion, respectively. Therefore, this map is similar to a frontal view of the human body. Colors indicate the manually labeled lesion types. Sample lesions are exhibited to show the great diversity of DeepLesion, including: a. lung nodule; b. lung cyst; c. costophrenic sulcus (lung) mass/fluid; d. breast mass; e.

liver lesion; f. renal mass; g. large abdominal mass; h. posterior thigh mass; i. iliac sclerotic lesion; j. perirectal lymph node (LN); k. pelvic mass; l. periportal LN; m. omental mass; n. peripancreatic lesion; o. splenic lesion; p. subcutaneous/skin nodule; q. ground glass opacity; r. axillary LN; s. subcarinal LN; t. vertebral body metastasis; u. thyroid nodule; v. neck mass.

Acknowledgments

This research was supported by the Intramural Research Program of the NIH Clinical Center. We thank NVIDIA for the donation of GPU cards. We thank our lab members Jiamin Liu, Yuxing Tang, and Youbao Tang for their help in preparing the dataset.

Questions & Comments

Ke Yan: ke.yan@nih.gov; Ronald Summers: rms@nih.gov