FAQ (Frequently Asked Questions):

Q01: I keep getting errors when I uncompressed the tar.gz files. Could you double check the integrity of files?

A: Those errors may be due to the incompleteness of downloaded files. Try to download the files one by one from different computers and places.

Q02: Are the radiological reports included in the publicly accessible chest X-ray dataset?

A: No. We currently do not have a plan to release the radiological reports.

Q03: Is the source code or pre-trained model available for the unified classification and localization framework introduced in the paper?

A: Currently, we do not plan to release the code or pre-trained model.

Q04: Are there any restrictions in using this dataset?

A: The usage of the data set is unrestricted. But you should provide the link to our original download site, acknowledge the NIH Clinical Center and provide a citation to our CVPR 2017 paper.

Q05: What does 'no finding' label mean?

A: 'No finding' means the 14 listed disease patterns are not found in the image.

Q06: Could you have the X-ray images available in DICOM format?

A: We are only able to provide the png images.

Q07: Do you have MD5s for the compressed files?

A: We thank Utku Ozbulak for providing those MD5s.

fe8ed0a6961412fddcbb3603c11b3698 images_001.tar.gz

ab07a2d7cbe6f65ddd97b4ed7bde10bf images_002.tar.gz

2301d03bde4c246388bad3876965d574 images_003.tar.gz

9f1b7f5aae01b13f4bc8e2c44a4b8ef6 images_004.tar.gz

1861f3cd0ef7734df8104f2b0309023b images_005.tar.gz

456b53a8b351afd92a35bc41444c58c8 images_006.tar.gz

1075121ea20a137b87f290d6a4a5965e images_007.tar.gz

b61f34cec3aa69f295fbb593cbd9d443 images_008.tar.gz

442a3caa61ae9b64e61c561294d1e183 images_009.tar.gz

09ec81c4c31e32858ad8cf965c494b74 images_010.tar.gz

499aefc67207a5a97692424cf5dbeed5 images_011.tar.gz

dc9fda1757c2de0032b63347a7d2895c images_012.tar.gz

Q08: General concerns about the image label accuracy.

A: There are several things about the published image labels that we want to clarify:

1. *Different terms and phrases might be used for the same finding: The image labels are mined from the radiology reports using NLP techniques. Those disease keywords are purely extracted from the reports. The radiologists often described the findings and impressions by using their own preferred terms and phrases for each particular disease pattern or a group of patterns, where the chance of using all possible terms in the description is small.*
2. *Which terms should be used: We understand it is hard if not impossible to distinguish certain pathologies solely based on the findings in the images. However, other information from multiple sources may be also available to the radiologists (e.g. reason for exam, patients' previous studies and other clinical information) when he/she reads the study. The diagnostic terms used in the report (like 'pneumonia') come from a decision based on all of the available information, not just the imaging findings.*
3. *Entity extraction using NLP is not perfect: we try to maximize the recall of finding accurate disease findings by eliminating all possible negations and uncertainties of disease mentions. Terms like 'It is hard to exclude …' will be treated as uncertainty cases and then the image will be labeled as 'No finding'.*
4. *'No finding' is not equal to 'normal'. Images labeled with 'No finding' could contain disease patterns other than the listed 14 or uncertain findings within the 14 categories.*
5. *We encourage others to share their own labels, ideally from a group of radiologists so that observer variability can also be assessed. The published image labels are a first step at enabling other researchers to start looking at the problem of 'automated reading a chest X-ray' on a very large dataset, and the labels are meant to be improved by the community.*

Q09: Will you publish the data split files?

A: Yes, two split files (train_val_list.txt and test_list.txt) are now provided and ready for downloading. Images in the ChestX-ray dataset are divided into two sets on the patient level. All studies from the same patient will only appear in either training/validation or testing set. This data split is adopted to generate the classification and localization results reported in our latest arxiv paper (ARXIV_V5_CHESTXRAY.pdf).