1 Datasheet for Dataset

1.1 Motivation

For what purpose was the dataset created?
The proposed RGC dataset is created for medical vision-language pre-training and to serve as a benchmark for medical image-text retrieval and report generation.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
The dataset is created by the research team led by Dr Xiao-Ming Wu at the Department of Computing, The Hong Kong Polytechnic University. The data is collected from and hosted on the MedPix[1] website.

1.2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?
Each instance in the dataset represents a radiographic image and its caption.

How many instances are there in total (of each type, if appropriate)?
There are 18,434 instances in total.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?
The dataset is a subset of the cases on MedPix website.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features?
"Raw" data.

Is there a label or target associated with each instance?
Yes. Each instance is an image-caption pair.

Are there recommended data splits (e.g., training, development/validation, testing)?
The dataset is split into a training and test set with a ratio of 9:1.

Are there any errors, sources of noise, or redundancies in the dataset?
It is possible that the noise was not completely removed by manual filtering.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?
Each instance of the dataset contains a URL that links to a medical case (an image-caption pair) on MedPix website, since many cases on MedPix cannot be redistributed.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)?
No.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

[1] `https://medpix.nlm.nih.gov/`

No.

Does the dataset relate to people?

Yes. All the cases on MedPix are collected from patients.

Does the dataset identify any subpopulations?

No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?

No.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?

No.

1.3 Collection Process

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The dataset was manually filtered from the MedPix database.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors)?

The research students Li Xu and Bo Liu (the first two authors of the paper) collected the dataset.

Were any ethical review processes conducted (e.g., by an institutional review board)?

No need. The data was collected from an online open-access database MedPix.

1.4 Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?

Yes. As described in the main paper, we attempted to remove noisy cases.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?

Yes. All the raw data is kept on the MedPix database.

Is the software that was used to preprocess/clean/label the data available?

No. Most cases were filtered manually. But if some errors (e.g., typos or noisy image-caption pairs) are found, we can provide a Python script to fix them.

1.5 Uses

Has the dataset been used for any tasks already?
Yes. We used the dataset to pre-train a medical vision-language Transformer, and as a benchmark to evaluate state-of-the-art models for report generation and medical image-text retrieval. The results are reported in the paper.
What (other) tasks could the dataset be used for?
In additional to vision-language pre-training, report generation, and medical image-text retrieval, the dataset can be potentially used for text-guided visual feature learning.

1.6 Distribution

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?
The dataset has already been hosted on MedPix: `https://openi.nlm.nih.gov/imgs/collections/RGC.zip`. Each instance of the dataset contains a URL that links to a medical case (an image-caption pair) on the MedPix website. Note that many cases on MedPix cannot be redistributed due to copyright issues.
Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?
Yes. The dataset is under MedPix License[2] and National Library of Medicine (NLM) Copyright[3].

1.7 Maintenance

Who will be supporting/hosting/maintaining the dataset?
The dataset has been hosted on the MedPix website. The authors will coordinate with MedPix team to provide support to the dataset. For any queries about the dataset, please contact the first author (`li-control.xu@connect.polyu.hk`) or the last author (`xiao-ming.wu@polyu.edu.hk`) of the paper.
How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
The authors can be reached by emails provided in the paper. The MedPix team can be reached by emails provided on MedPix website.
Is there an erratum?
No.
Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?
Yes.
Will older versions of the dataset continue to be supported/hosted/maintained?
Yes.

[2] `https://medpix.nlm.nih.gov/licensing`

[3] `https://www.nlm.nih.gov/web_policies.html#copyright`

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

Yes. They may contact the authors or MedPix team.