DATASET DESCRIPTION

VQA-MED 2018 dataset

The medical VQA-MED dataset developed from radiology images along with their captions from PubMed Central articles and part of the ImageCLEF 2017 caption prediction task. The two expert human annotators were assigned to check and validate the question-answer pairs associated with the radiology images in two passes. The syntactic and semantic correctness were checked by a human annotator in the first pass. In the second pass, the validation and test sets were evaluated for clinical relevance by a medical expert, the second annotator. As a result, the corpus consists of 2,866 medical images associated with 6,413 question-answer pairs that have been finalized.

VQA-MED 2019 dataset

The VQA-MED 2019 dataset shows an increase in the number of samples and their diversity, and it is more organized than the 2018 dataset. This is achieved by selecting relevant radiology images from the MedPix database with filters corresponding to captions, planes, categories, modalities, localities, and diagnostic methods. The VQA-MED 2019 dataset consists of samples related to sixteen planes, ten organs, and thirty-six modalities. As a result, the overall dataset size has increased to 4,200 images, each associated with 3 to 4 questions, along with 12,792 QA pairs. Further, a medical doctor and a radiologist performed a manual double validation of the test answers and corrected ten questions, which is 3% of the total test set size.

VQA-MED 2020 dataset

The VQA-MED 2020 dataset focuses on abnormality-type samples of different organs, planes, and modalities. The ImageCLEF forum created this dataset automatically through the process of (i). Applying filters to select relevant images and associated annotations (ii). Creating patterns to generate questions and answers (iii) Selecting relevant medical images from the Med-Pix database based on their captions, localities, and diagnostic methods. The final list has 330 medical problems, where each problem occurs at least 10 times in the created VQA data. This dataset consists of 5000 radiology images and 5000 QA pairs that are divided among training, validation, and test sets. The most common medical problems and their frequencies in this dataset are pulmonary embolism (114), acute appendicitis (109), angiomyolipoma (68), osteochondral (63), adenocarcinoma of the lung (60), and sarcoidosis (58).

VQA-MED 2021 dataset

The VQA-MED 2021 dataset is primarily on abnormality-type questions for different categories like organs, planes, and modalities, like the previous year. This dataset consists of 5,500 samples, of which 4,500 are from the VQA-MED 2020 dataset, and the complete dataset is divided into training, validation, and test sets. Compared to 2020, the quality of the 2021 data has shown a marked improvement by having the reference answers of the test set validated by a medical doctor.

Path-VQA dataset

The Path-VQA dataset consists of pathological VQA queries collected from online digital libraries and pathology textbooks. The pathology images in the Path-VQA dataset are acquired from the generic textbook named “Textbook of Pathology” by semi-automated pipeline approach. The image and the QA pairs are mapped sequentially, and the annotations are reviewed by the medical annotator.