Visual Dialog for Radiology: Data Curation and First Steps

Olga Kovaleva[*†] okovalev@cs.uml.edu
Chaitanya Shivade[*§] shivadc@amazon.com
Satyananda Kashyap [¶] satyananda.kashyap@ibm.com

Karina Kanjaria [¶] **Adam Coy** [¶] **Deddeh Ballah** [¶] **Yufan Guo** [¶] **Joy Wu** [¶]

Alexandros Karargyris [¶] **David Beymer** [¶] **Anna Rumshisky** [‡] **Vandana Mukherjee** [¶]

Abstract

Recent work in clinical AI has been focusing on solving tasks that involve both image understanding and reading comprehension. In this study, we further pursue this line of research and introduce the first Visual Dialog task in Radiology, which adds complexity to existing tasks. We present our data collection strategy for both silver and gold-standard datasets for chest x-ray images and discuss associated challenges. We evaluate a Stacked Attention Network model, commonly used for Visual Question answering in medical domain, and provide baseline results indicating the difficulty of the task.

1 Introduction

Answering questions about an image is a complex multi-modal task demonstrating an important capability of artificial intelligence. A well-defined task evaluating such capabilities is Visual Question Answering (VQA) Antol et al. [2015] where a system answers free-form questions reasoning about an image. VQA requires understanding of the intricacies of both the image and the language used in framing the question. Visual Dialog (VisDial) Das et al. [2017], de Vries et al. [2016] is an extension to the VQA problem where a system is required to engage in a dialog about the image. This adds significant complexity to VQA where a system should now be able to ground the question in the image, and reason over additional information gathered from previous question answers in the dialog.

Unlike other domains, the medical domain poses a unique set of problems in terms of data availability and annotations. The mandate of patient privacy restricts the sharing of data for research. However, the research community has acknowledged this limitation and publicly available de-identified medical image data Wang et al. [2017], Demner-Fushman et al. [2015], Irvin et al. [2019] has been made available at least with limited labels. Further, such data is intelligible only to experts skilled with the necessary domain knowledge. Therefore, a standard pipeline of collecting crowd-sourced annotations followed by training of deep neural networks is expensive and often not possible.

However, most of the current research in machine learning for radiology discards any other information about the patient and is focused entirely on images Litjens et al. [2017]. To this end, we explore the problem of visual dialog in radiology specific to chest X-ray images. Answering questions about

[*] Authors have equal contribution
[†] University of Massachusetts Lowell
[‡] Work done at IBM Research
[§] Amazon
[¶] IBM Research

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

a radiology image is a challenging task. Reasoning over the medical history and a dialog adds further linguistic complexity. The medical domain has a distinct sub-language with its own vocabulary differentiating it from the open domain. VisDial naturally fits in the workflow of a radiology read where answering questions about an image should become easier as more information is known about the patient. Although limited work exploring VQA in radiology exists, VisDial in radiology remains an unexplored problem.

In this paper, we introduce and make publicly available the first data set for visual dialog in radiology derived from MIMIC-CXR Johnson et al. [2019] called RadVisDial. We detail the construction of a silver standard dataset purely from parsing the available MIMC-CXR reports and report baseline results using a stacked attention network Yang et al. [2016]. Finally, we detail the steps underway and the challenges to collect a true gold standard visual dialog data set between two doctors on the MIMIC-CXR dataset.

2 Related Work

Most of the large publicly available datasets Kaggle [2017], Rajpurkar et al. [2017] for radiology consist of images with limited amount of structured information associated with them. For example, Irvin et al. [2019], Johnson et al. [2019] make images available along with the output of a text extraction module that produces labels for 13 abnormalities in a chest X-ray. Two recent shared tasks at ImageCLEF explored the VQA problem with radiology images Hasan et al. [2018], Abacha et al. [2019]. Lau et al. [2018] also released a small dataset VQA-RAD for the specific task. The first VQA shared task at ImageCLEF Hasan et al. [2018] used images from articles at PubMed Central. While Abacha et al. [2019] and Lau et al. [2018] use clinical images, the sizes of these datasets are limited. They are a mix of several modalities including two dimensional modalities such as x-rays and three dimensional modalities such as ultrasound, MRI, and CT scans. They also cover several anatomies from the brain to the limbs. This makes a multi-modal task with such images overly challenging with shared task participants developing separate models Al-Sadi et al. [2019], Abacha et al. [2018], Kornuta et al. [2019] to first address these problems before actually solving the problem of VQA. Table 1 presents the summary of existing tasks and their differences.

Dataset	Task	Modality	# ques.	# images	Image source
Kaggle [2017]	Image classification	CT	-	1K	National Cancer Institute
VQA-Med-2018	VQA	Multiple	6K	3K	PubMed Central articles
VQA-Med-2019	VQA	Multiple	15K	4K	MedPix database
VQA-RAD	VQA	Multiple	3.5K	315	MedPix database
CheXpert	Image classification	X-ray	-	225K	Stanford Hospital
VisDial (ours)	Visual Dialog	X-ray	450K	91K	MIMIC-CXR

Table 1: Comparison of existing tasks and datasets.

3 MIMIC-CXR Data

The MIMIC-CXR dataset[6] consists of 371,920 chest X-ray images in the Digital Imaging and Communications (DICOM) format along with 206,576 reports. Each report is well structured and typically consists of sections such as `Medical Condition`, `Comparison`, `Findings`, and `Impression`. Each report can map to one or more image and each patient can have one or more reports. The dataset corresponds to radiographic studies conducted at the Beth Israel Deaconess Medical Center in Boston, MA, USA and is de-identified to meet Health Insurance Portability and Accountability Act (HIPAA) requirements. The images consist of both frontal (either anterior-posterior (AP) or posterior-anterior (PA)) and lateral views. The initial release of data also consists of annotations for 14 labels (13 abnormalities and one `No Findings` label) for each image. These annotations are obtained by running the CheXpert labeler Irvin et al. [2019]; a rule based NLP

[6] `https://physionet.org/content/mimic-cxr/1.0.0/`

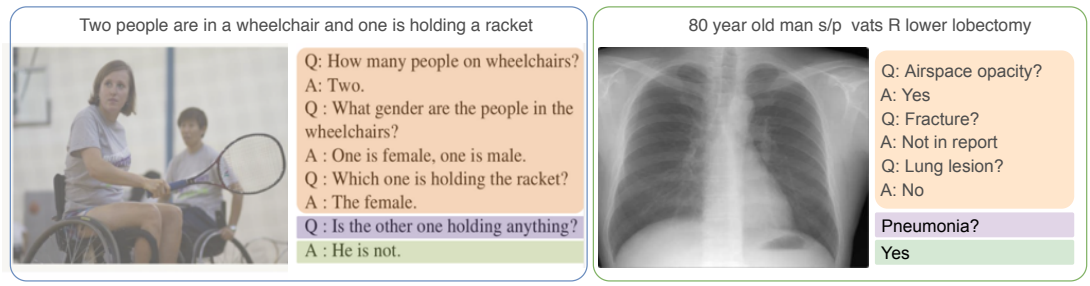

Figure 1: Comparison of VisDial 1.0 (left) with our synthetically constructed dataset (right).

pipeline against the associated report. The labeler output indicates one of four possibilities for each of the 13 abnormalities: $\{$*yes, no, maybe, not mentioned in the report*$\}$.

4 Silver Standard Dataset

4.1 Data Creation

Every training record of the original VisDial dataset consists of three elements: an image I, a caption for the image C, and a dialog history H consisting of a sequence of ten question-answer pairs. The task is given I, C, a possibly empty dialog history H, and a follow-up question q, to generate an answer a where $\{q, a\} \in H$. We synthetically create our dataset using the plain text reports associated with each image. The `Medical Condition` section of the radiology report is a single sentence describing the medical history of the patient. We use NegBio Peng et al. [2018] for extracting sections within a report and treat this sentence from the `Medical Condition` section as the *caption* of the image. We discard all the images that did not have a medical condition in their report. Further, each CheXpert label is formulated as a question probing the presence of a disorder and the output from the labeler is treated as the corresponding answer. Thus, ignoring the `No Findings` label, there are 52 possible question-answer pairs as a result of 13 questions and 4 possible answers.

Our radiologists recommended starting on PA images for most of our experiments since this is the most informative view for chest X-rays. Since we had only 13 questions, we limited the length of the dialogs to 5 randomly sampled questions from the set of 13 questions. The resulting dataset has 91060 images in the PA view. This synthetic data will be made available through the MIMIC Derived Data Repository[7]. Thus any individual with access to MIMIC-CXR will have access to our data. Figure 1 shows an example training record from our dataset and how it compares with one from VisDial 1.0 (Das et al. [2017]).

4.2 Model Description

For our experiment we used the Stacked Attention Network (SAN) Yang et al. [2016] model. Following the original Visual Dialog study Das et al. [2017], we use an encoder-decoder structure with a discriminative decoder for each of the models. During training, we set the hidden dimensionality of the used LSTMs to 512, and the learning rate to $2 \cdot 10^{-4}$. We use Adam optimizer and a batch size of 256.

The original configuration of SAN was introduced for the general-domain Visual Question Answering task. The model performs multi-step reasoning by refining question-guided attention over image features in an iterative manner. The attended image features are then combined with the question features for answer prediction. SAN has been successfully adapted for medical VQA tasks such as VQA-RAD Lau et al. [2018] and VQA-Med task of the ImageCLEF 2018 challenge Ionescu et al. [2018]. In our setup we use a stack of two image attention layers and an LSTM-based question representation.

To take the dialog history into account and therefore adjust the SAN model for the needs of the Visual Dialog task, we modify the first image attention layer of the network by adding a term for LSTM

[7] `https://physionet.org/physiotools/mimic-code/HEADER.shtml`

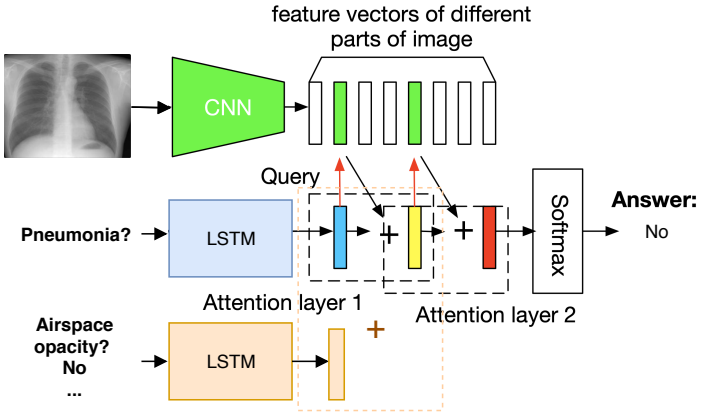

Figure 2: The modified architecture of the SAN model (image taken from Yang et al. [2016]). The proposed modification shown in orange incorporates the history of dialog turns in the same way as the question through an LSTM. In our ablation experiments the changed part either reduces to encoding an image caption only or gets cut completely.

representation of the history. This modification forces the image attention weights to become both question- and history-guided (see Figure 2).

4.3 Evaluation

Questions in our dataset are limited to probe the presence of an abnormality in the chest X-ray. Similarly, the answers are limited to one of the four choices. Owing to the confined nature of the problem, we deviate from the evaluation protocol outlined in Das et al. [2017] and reported a macro F1 score of 0.243.

4.4 Planned Experiments

Given the poor results, we plan to investigate in the following concurrent directions: 1) exploring different state of the art networks for visual dialog, 2) data balancing - our initial analysis showed that most of the answers were *Not in report* heavily skewing the distribution; therefore, it will be beneficial to explore the balancing strategy used in Hudson and Manning [2019], 3) combining lateral and frontal views of chest X-rays - doctors usually benefit from an additional view for certain set of diseases. 4) experimenting with various textual and image representations. Our plans for future work also include rephrasing the questions (i.e. CheXpert labels) in a more natural language form. While paraphrasing can be easily done with the help of templates and external sources such as UMLSBodenreider [2004], it will make the task more challenging and bring it closer to real-world applications.

5 Gold Standard Dataset

Our work in visual dialog started off with a silver standard dataset making using of the CheXpert parser to generate a dialog. However, we recognize that expert dialog on a chest X-ray image is necessary to truly see the benefits of visual dialog in radiology. We have started the process of collecting real dialog between two expert radiologists. There are several factors to be considered as listed below:

- **The cost of expert annotators**. The primary bottleneck in getting gold standard annotations in radiology is the high cost of expert radiologists. Visual dialog for radiology further requires two radiologists to have a conversation increase the annotation cost two fold. Given the constraints, we were not able to ask radiologists for additional useful annotations such as image bounding boxes for a given pathology.
- **The dialog medium and workflow**. One major issue with medical data is to run it through HIPAA compliant medium. We are exploring optimal methods wherein the doctors can have

a dialog and also reduce the overhead in terms of loading images, meta-data and saving the resulting dialogues

- **Minimum dialog length**. This heuristic is being explored to determine the optimal number of dialog rounds. Based on consultation with our radiologists, we initially set the minimum number of dialog turns to 5 for our silver-standard data. A pilot study for the gold-standard data is underway to make necessary changes, if needed.
- **Single or multiple view images**. Given how the radiologists advised us to use PA frontal view images for the silver-standard dataset, we are now exploring whether providing multiple-view images could simulate their real radiology reads.

6 Discussion

This paper discussed the efforts undertaken in developing visual dialog for radiology. The method used to curate a silver standard dataset for visual dialog in radiology was outlined. Further we discussed concurrent steps that will be taken to address some of the shortcomings we encountered in our initial study. Finally we outlined the factors that we needed to address to collect the gold standard data. Efforts are underway to collect dialogs between two expert radiologists on the MIMIC chest X-rays.

References

AB Abacha, SA Hasan, VV Datla, J Liu, D Demner-Fushman, and H Müller. Vqa-med: Overview of the medical visual question answering task at image-clef 2019. In CLEF2019 Working Notes. CEUR Workshop Proceedings (CEURWS. org), ISSN, pages 1613–0073, 2019.

Asma Ben Abacha, Soumya Gayen, Jason J Lau, Sivaramakrishnan Rajaraman, and Dina Demner-Fushman. Nlm at imageclef 2018 visual question answering in the medical domain. In CLEF (Working Notes), 2018.

Aisha Al-Sadi, Bashar Talafha, Mahmoud Al-Ayyoub, Yaser Jararweh, and Fumie Costen. Just at imageclef 2019 visual question answering in the medical domain. In CLEF (Working Notes), 2019.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqaa: Visual question answering. In Proceedings of the IEEE International Conference on Computer Vision, pages 2425–2433, 2015.

Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research, 32(suppl_1):D267–D270, 2004.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 326–335, 2017.

Harm de Vries, Florian Strub, A. P. Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron C. Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In 2017 IEEE Conference on Computer Vision and Pattern Recognition, pages 4466–4475, 2016.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association, 23(2):304–310, 2015.

Sadid A Hasan, Yuan Ling, Oladimeji Farri, Joey Liu, Henning Müller, and Matthew Lungren. Overview of imageclef 2018 medical domain visual question answering task. In CLEF (Working Notes), 2018.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6700–6709, 2019.

Bogdan Ionescu, Henning Müller, Mauricio Villegas, Alba García Seco de Herrera, Carsten Eickhoff, Vincent Andrearczyk, Yashin Dicente Cid, Vitali Liauchuk, Vassili Kovalev, Sadid A Hasan, et al. Overview of imageclef 2018: Challenges, datasets and evaluation. In International Conference of the Cross-Language Evaluation Forum for European Languages, pages 309–334. Springer, 2018.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of AAAI, 2019.

Alistair EW Johnson, Tom J Pollard, Seth Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr: A large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042, 2019.

Kaggle. Data science bowl. `https://www.kaggle.com/c/data-science-bowl-2017`, 2017.

Tomasz Kornuta, Deepta Rajan, Chaitanya Shivade, Alexis Asseman, and Ahmet S Ozcan. Leveraging medical visual question answering with supporting facts. In CLEF (Working Notes), 2019.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. Scientific Data, 5:180251, 2018.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. Medical Image Analysis, 42:60–88, 2017.

Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald Summers, and Zhiyong Lu. Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. AMIA Summits on Translational Science Proceedings, 2018:188, 2018.

Pranav Rajpurkar, Jeremy Irvin, Aarti Bagul, Daisy Ding, Tony Duan, Hershel Mehta, Brandon Yang, Kaylie Zhu, Dillon Laird, Robyn L Ball, et al. Mura: Large dataset for abnormality detection in musculoskeletal radiographs. arXiv preprint arXiv:1712.06957, 2017.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2097–2106, 2017.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 21–29, 2016.