Towards Visual Dialog for Radiology

Olga Kovaleva‡,* Chaitanya Shivade†*, Satyananda Kashyap⋆, Karina Kanjaria⋆, Adam Coy⋆, Deddeh Ballah⋆, Joy Wu⋆, Yufan Guo⋆, Alexandros Karargyris⋆, David Beymer⋆, Anna Rumshisky‡, Vandana Mukherjee⋆
‡ University of Massachussets, Lowell † Amazon
⋆ IBM Almaden Research Center.

Abstract

Current research in machine learning for radiology is focused mostly on images. There exists limited work in investigating intelligent interactive systems for radiology. To address this limitation, we introduce a realistic and information-rich task of Visual Dialog in radiology, specific to chest X-ray images. Using MIMIC-CXR, an openly available database of chest X-ray images, we construct both a synthetic and a real-world dataset and provide baseline scores achieved by state-of-the-art models. We show that incorporating medical history of the patient leads to better performance in answering questions as opposed to conventional visual question answering model which looks only at the image. While our experiments show promising results, they indicate that the task is extremely challenging with significant scope for improvement. We make both the datasets (synthetic and gold standard) and the associated code publicly available to the research community.

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 demands careful understanding of elements in an image along with intricacies of the language used in framing a question about it. 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 associate the question in the image, and reason over additional information gathered from previous question answers in the dialog.

Although limited work exploring VQA in radiology exists, VisDial in radiology remains an unexplored problem. With the healthcare setting increasingly requiring efficiency, evaluation of physicians is now based on both the quality and the timeliness of patient care. Clinicians often depend on official reports of imaging exam findings from radiologists to determine the appropriate next step. However, radiologists generally have a long queue of imaging studies to interpret and report, causing subsequent delay in patient care (Bhargavan et al., 2009; Siewert et al., 2016). Furthermore, it is common practice for clinicians to call radiologists asking follow-up questions on the official reporting, leading to further inefficiencies and disruptions in the workflow (Mangano et al., 2014).

Visual dialog is a useful imaging adjunct that can help expedite patient care. It can potentially answer a physician's questions regarding official interpretations without interrupting the radiologist's workflow, allowing the radiologist to concentrate their efforts on interpreting more studies in a timely manner. Additionally, visual dialog could provide clinicians with a preliminary radiology exam interpretation prior to receiving the formal dictation from the radiologist. Clinicians could use the information to start planning patient care and decrease the time from the completion of the radiology exam to subsequent medical management (Halsted and Froehle, 2008).

In this paper, we address these gaps and make the following contributions: 1) we introduce construction of RadVisDial - the first publicly available dataset for visual dialog in radiology, derived from the MIMIC-CXR (Johnson et al., 2019) dataset, 2) we compare several state-of-the-art models for VQA and VisDial applied to these images, and 3) we conduct a comprehensive set of experiments

* Equal contribution, Work done at IBM Research

Proceedings of the BioNLP 2020 workshop, pages 60–69
Online, July 9, 2020

highlighting different challenges of the problem and propose solutions to overcome them.

2 Related Work

Most of the large publicly available datasets (Kaggle, 2017; Rajpurkar et al., 2017) for radiology consist of images associated with a limited amount of structured information. 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. Of note recently, the task of generating reports from radiology images has become popular in the research community (Jing et al., 2018; Wang et al., 2018). 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 2D modalities such as X-rays, and 3D modalities such as ultrasound, MRI, and CT scans. They also cover several anatomic locations from the brain to the limbs. This makes a multimodal 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 subtasks (such as modality detection) before actually solving the problem of VQA.

We address these limitations and build up on MIMIC-CXR (Johnson et al., 2019) the largest publicly available dataset of chest X-rays and corresponding reports. We focus on the problem of visual dialog for a single modality and anatomy in the form of 2D chest X-rays. We restrict the number of questions and generate answers for them automatically which allows us to report results on a large set of images.

3 Data

3.1 MIMIC-CXR

The MIMIC-CXR dataset[1] 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 images and each patient can have one or more reports. The images consist of both frontal and lateral views. The frontal views are either anterior-posterior (AP) or posterior-anterior (PA). 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 pipeline against the associated report. The labeler output assigns one of four possibilities for each of the 13 abnormalities: {*yes, no, maybe, not mentioned in the report*}.

3.2 Visual Dialog dataset construction

Every training record of the original VisDial dataset (Das et al., 2017) 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. Given the image I, the caption C, a possibly empty dialog history H, and a follow-up question q, the task is to generate an answer a where $\{q, a\} \in H$. Following the original formulation, we synthetically create our dataset using the plain text reports associated with each image (this synthetic dataset will be considered to be silver-standard data for the experiments described in section 5). The `Medical Condition` section of the radiology report is a single sentence describing the medical history of the patient. We treat this sentence from the `Medical Condition` section as the *caption* of the image. We use NegBio (Peng et al., 2018) for extracting sections within a report.

We discard all images that do 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.

We decided to focus on PA images for most of our experiments as this is the most informative view for chest X-rays, according to our team radiologists. The original VisDial dataset (Das et al., 2017) consists of ten questions per dialog and one dialog per image. Since we only have a set of 13

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

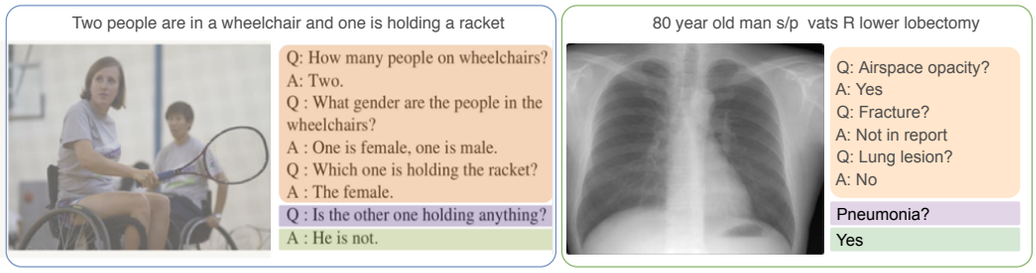

Figure 1: Comparison of VisDial 1.0 (left) with our synthetically constructed dataset (right).

possible questions, we limit the length of the dialog to 5 randomly sampled questions. The resulting dataset has 91060 images in the PA view (with train/validation/test splits containing 77205, 7340 and 6515 images, respectively). This synthetic data will be made available through the MIMIC Derived Data Repository.[2] Thus any individual with access to MIMIC-CXR will have access to our data. Figure 1 shows an example from our dataset and how it compares with one from VisDial 1.0.

3.3 Evaluation

The questions in our dataset are limited to probing the presence of an abnormality in a chest X-ray. Similarly, the answers are limited to one of the four choices. Owing to the restricted nature of the problem, we deviate from the evaluation protocol outlined in (Das et al., 2017) and instead calculate the F1-score for each of the four answers. We also report a macro-averaged F1 score across the four answers to make model comparisons easier.

4 Models

For our experiments, we selected a set of models designed for image-based question answering tasks. Namely, we experimented with three architectures: Stacked Attention Network (SAN) (Yang et al., 2016), Late Fusion Network (LF) (Das et al., 2017), and Recursive Visual Attention Network (RVA) (Niu et al., 2019). Following the original VisDial study (Das et al., 2017), we use an encoder-decoder structure with a discriminative decoder for each of the models. Below we give an overview of all the three algorithms.

4.1 Stacked Attention Network

The original configuration of SAN was introduced for the general-domain VQA 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.

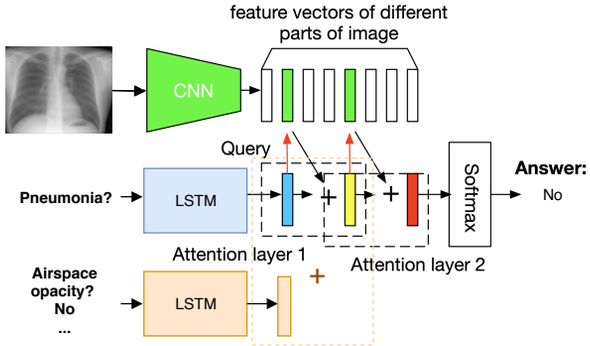

Figure 2: The modified architecture of the SAN model (image taken from (Yang et al., 2016)). The proposed modification shown in orange incorporates the history of dialog turns in the same way as the question through an LSTM. In our ablation experiments the changed part either reduces to encoding an image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

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

4.2 Late Fusion Network

Proposed by (Das et al., 2017) as a baseline model for the Visual Dialog task, Late Fusion Network encodes the question and the dialog history through two separate RNNs, and the image through a CNN. The resulting representations are simply concate-

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

nated in a single vector, which is then used by a decoder for predicting the answer. We use this model unchanged, as released in the original Visual Dialog challenge.

4.3 Recursive Visual Attention

This model is the winner of the 2019 Visual Dialog challenge[3]. It recursively browses the past history of dialog turns until the current question is paired with the turn containing the most relevant information. This strategy is particularly useful for resolving co-references, naturally occurring in general-domain dialog questions. As previously, we do not modify the architecture of the model.

5 Experiments

This section presents our down-sampling strategy, gives details about conducted ablation studies, and describes experiments with various representations of images and texts.

5.1 Downsampling

A closer analysis of our data showed that the majority of the reports processed by the CheXpert labeler resulted in no mention of most of the 13 pathologies. This presented a heavily skewed dataset that would lead to a biased model instead of true visual understanding. This issue is not unique to radiology; it is observed even in the current benchmarks for VQA, and attempts have been made to mitigate the resulting problems (Hudson and Manning, 2019; Zhang et al., 2016; Agrawal et al., 2018).

In order to dissuade the answer biases, we performed data balancing, specifically by downsampling major labels in our dataset. As mentioned above, the CheXpert labeler outputs four possible answers for 13 labels. To investigate the skew in the data, we plotted a distribution of the 52 question-answer pairs (Figure 3). Further, we downsampled the question-answer pairs to fit a smoother answer distribution with the method presented in GQA based on the Earth Mover's Distance method (Hudson and Manning, 2019; Rubner et al., 2000). We iterated over the 52 pairs in decreasing frequency order and downsampled the categories belonging to the skewed head of the distribution. The relative label ranks by frequency remained the same for the balanced sets as with the unbalanced sets. For example, the pairs {*'Other pleural findings'* → *'Not in report'* } and {*'Fracture'* → *'Not in report'* } remained the first and second largest counts in both the unbalanced and downsampled versions of the datasets. To reduce the disparity between dominant and underrepresented categories, we tuned the parameters outlined in (Hudson and Manning, 2019). We experimented with two different sets of parameter values and obtained two datasets with more balanced question-answer distributions. We further refer to them as "minor" and "major" downsampling, reflecting the total amount of data reduced (shown in blue and gray in Figure 3).

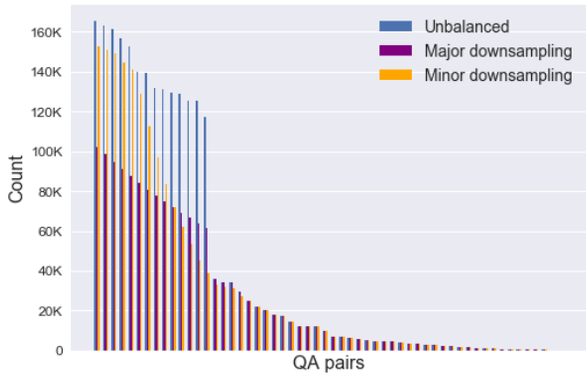

Figure 3: Downsampling strategies. Every bar along the X axis represents a single question-answer pair, where questions (13 in total) and answers (4 in total) are obtained through CheXpert.

5.2 Evaluating importance of context

To assess the importance of the dialog context for question answering, we compare the performance of different variations of the Stacked Attention Network, selected as the best-performing model in the previous experiment (see subsection 6.1). In particular, we examine three scenarios: (a) the model makes a prediction based solely on a given image (essentially solving the VQA task rather than the Visual Dialog task), (b) the model makes its prediction given an image and its caption, and (c) the model makes its prediction given an image, a caption, and a history of question-answer pairs. Similar to the model modifications described in subsection 4.1 and Figure 2, we achieve the goal through experimenting with the SAN model by changing its first image attention layer to accordingly take in (a) question and image features, (b) question, image, and caption features, and (c) question, image, and full dialog history features.

[3] https://visualdialog.org/challenge/2019

5.3 Image representations

We test three approaches for pre-trained image representations. The first approach uses a ResNet-101 architecture (He et al., 2016) for multiclass classification of input X-ray images into 14 finding labels extracted from the associated reports (as described in section 3.2). Our second method aims to replicate the original CheXpert study (Irvin et al., 2019). Here we use a DenseNet-121 image classifier trained for prediction of five pre-selected and clinically important labels, namely, atelectasis, cardiomegaly, consolidation, edema, and pleural effusion. In both ResNet and DenseNet-based approaches we take the features obtained from the last pooling layer.

Finally, we adopted a bottom-up mechanism for image region proposal introduced by Anderson et al. (2018). More specifically, we first trained a neural network predicting bounding boxes for the image regions, corresponding to a set of 11 handcrafted clinical annotations adopted from an existing chest X-ray dataset[4]. We then represented every region as a latent feature vector of a trained patch-wise convolution autoencoder, and (3) concatenated all the obtained vectors to represent the entire image.

Based on the results of the experiment (subsection 6.3), we found that ResNet-101 image vectors yielded the best performance, so we used them in other experiments.

5.4 Effect of incorporating a lateral view

One of the crucial aspects of X-ray radiography exams is to capture the subject from multiple views. Typically, in case of chest X-rays, radiologists order an additional lateral view to confirm and locate findings that are not clearly visible from a frontal (PA or AP) view. We test whether the VisDial models are able to leverage the additional visual information offered by a lateral (LAT) view. We filter the data down to the patients whose chest X-ray exams had both a frontal and lateral views and re-sample the resulting data-set into train (52952 PA and 8086 AP images), validation (6614 PA and 964 AP images), and test (6508 PA and 1035 AP images). We train a separate ResNet-101 model for each of the three views on this re-sampled data using the method described in the previous section. The vector representations of a frontal view and the corresponding lateral view are concatenated as an aggregate image representation.

[4] https://www.kaggle.com/c/rsna-pneumonia-detection-challenge

5.5 Text representations

Finally, we investigate the best way for representing the textual data by incorporating different pre-trained word vectors. More specifically, we measure the performance of our best-performing SAN model reached with (a) randomly initialized word embeddings trained jointly with the rest of the models, (b) domain-independent GloVe Common Crawl embeddings (Pennington et al., 2014), and (c) domain-specific fastText embeddings trained by (Romanov and Shivade, 2018). The latter are initialized with GloVe embeddings trained on Common Crawl, followed by training on 12M PubMed abstracts, and finally on 2M clinical notes from MIMIC-III database (Johnson et al., 2016). In all the experiments, we use 300-dimensional word vectors. We also experimented with transformer-based contextual vectors using BERT (Devlin et al., 2019). More specifically, instead of using LSTM representations of the textual data, we extracted the last layer vectors from ClinicalBERT (Alsentzer et al., 2019) pre-trained on MIMIC notes, and averaged them over input sequence tokens.

5.6 Question order

In a visual dialog setting, a model is conditioned on the image vector, the image caption, and the dialog history to predict the answer to a new question. We hypothesized that a model should be able to answer later questions in a dialog better since it has more information from the previous questions and their answers. As described in Section 3.2, we randomly sample 5 questions out of 13 possible choices to construct a dialog. We re-ordered the question-answer pairs in the dialog to reflect the order in which the corresponding abnormality label mentions occurred in the report. However, results for questions ordered based on their occurrence in the narrative did not vary from the setup with a random order of questions.

6 Results

We report macro-averaged F1-scores achieved on the same unbalanced validation set for each of the experiments. When experimenting with different configurations of the same model, we also break down the aggregate score to the F1 scores for individual answer options.

Model	'Yes'	'No'	'Maybe'	'Not in report'	Macro F1
SAN (VQA)	0.24	0	0.09	0.84	0.29
SAN (caption only)	0.30	0.09	0.09	0.81	0.33
SAN (full history)	0.22	0.26	0.04	0.83	**0.34**

Table 1: Ablation experiments. Per-answer F1-scores along with the macro F1-score are shown for tested SAN configurations.

6.1 Downsampling

Our results show (Table 2) consistent improvement of the scores across all the models as the training data becomes more balanced. All the models yielded comparable scores, with SAN being slightly better than other models (0.34 against 0.33 macro F1-score). Later in our experiments, we used the major down sampled version of the data-set.

Model	Unbalanced	Downsampled	
		Minor	Major
SAN	0.25	0.28	0.34
LF	0.28	0.31	0.33
RvA	0.24	0.33	0.33

Table 2: Data balancing experiments. Macro F1 scores are reported for every tested model.

6.2 Evaluating importance of context

One of the main findings of our study revealed the importance of contextual information for answering questions about a given image. As shown in Table 1, adding the image caption and the history of turns results in incremental increases of macro F1-scores. Notably, the VQA setup in which the model relies on the image only, it fails to detect the *'No'* answer, whereas the history-aware configuration leads to a significant performance gain for this particular label. As expected and due to the skewed nature of the data-set, the highest and the lowest per-label scores were achieved for the most and the least frequent labels (*'Not in report'* and *'Maybe'*), respectively.

6.3 Image representation

Out of the tested image representations, ResNet-derived vectors perform consistently better than the other approaches (see Table 3). Although in our DenseNet-121 image classification pre-training we were able to replicate the performance of (Irvin et al., 2019), the Visual Dialog scores for the corresponding vectors turned out to be lower. We believe this might be due to the fact that, by design, the network uses a limited set of pre-training classes not sufficient to generalize well to a full set of diseases used in the Visual Dialog task.

Model	DenseNet-121	Region Proposal	ResNet
SAN	0.27	0.29	**0.34**
LF	**0.33**	0.31	**0.33**
RvA	0.29	0.32	**0.33**

Table 3: Comparative performance (macro-F1) of Visual Dialog models on the test set with different image representations.

6.4 Effect of incorporating a lateral view

As expected, for both variations of the frontal view (i.e. AP and PA) appending lateral image vectors enhanced the performance of the tested SAN model (see Table 4). This suggests that lateral and frontal image vectors complement each other, and the models can benefit from using both. However, in our data-set only a subset of reports has both views available, which significantly reduces the amount of training data.

6.5 Word embeddings

Another observation from our experiments is that domain-specific pre-trained word embeddings contribute to better scores (see Table 5). This is due to the fact that domain-specific embeddings contain medical knowledge that helps the model make more justified predictions.

When using BERT, we did not notice gains in performance, which most likely means that the last-layer averaging strategy is not optimal and more sophisticated approaches such as (Xiao, 2018) are required . Alternatively, the final representation of the `CLS` can be used to represent input text.

View		'Yes'	'No'	'Maybe'	'Not in report'	Macro F1
AP+LAT	AP	0.40	0.21	0.12	0.79	0.381
	LAT	0.41	0.23	0.13	0.75	0.379
	AP + LAT	0.41	0.22	0.12	0.79	**0.385**
PA+LAT	PA	0.30	0.30	0.08	0.88	0.392
	LAT	0.32	0.32	0.07	0.86	0.391
	PA + LAT	0.32	0.34	0.06	0.87	**0.396**

Table 4: Effect of adding the lateral view to a frontal view (AP and PA).

Embedding	'Yes'	'No'	'Maybe'	'Not in report'	Macro F1
Random	0.26	0.22	0.04	0.73	0.31
GloVe (common crawl)	0.27	0	0.09	0.80	0.29
fastText (MedNLI)	0.24	0.22	0.07	0.84	**0.33**

Table 5: Comparative performance of the SAN model with different word embeddings.

7 Comparison with the gold-standard data

To complement our experiments with the silver data and investigate the applicability of the trained models to real-world scenarios, we also collected a set of gold standard data which consisted of two expert radiologists having a dialog about a particular chest X-ray. These X-ray images were randomly sampled PA views from the test our data. In this section, we present the data collection workflow, outline the associated challenges, compare the resulting data-set with the silver-standard, and report the performance of trained models.

7.1 Gold Standard Data Collection

We laid the foundations for our data collection in a manner similar to that of the general visual dialog challenge (Das et al., 2017). Two radiologists, designated as a "questioner" and an "answerer", conversed with each other following a detailed annotation guideline created to ensure consistency. The "answerer" in each scenario was provided with an image and a caption (medical condition). The "questioner" was provided with only the caption, and tasked with asking follow-up questions about the image, visible only to the "answerer". In order to make the gold data-set comparable to the silver-standard one, we restricted the beginning of each answer to contain a direct response of *'Yes'*, *'No'*, *'Maybe'*, or *'Not mentioned'*. In our annotation guidelines *'Not mentioned'* referred to the lack of evidence of the given medical condition that was asked by the "questioner" radiologist. The answer was elaborated with additional information if the radiologists found it necessary. The whole data collection procedure resulted in 100 annotated dialogs.

7.2 Gold standard results

Following the gold standard data collection, we performed some preliminary analyses with the best silver standard SAN model. Our gold standard data was split into train (70), validation (20), and test (10) sets. We experimented with three setups: (a) evaluating the silver-data trained networks on the gold standard data, (b) training and evaluating the models on the gold data, and (c) fine-tuning the silver-data trained networks on the gold standard data. Table 6 shows the results of these experiments. We found the best macro-F1 score of 0.47 was achieved by the silver data-trained SAN network fine-tuned on the gold standard data. We observed that the model could not directly predict any of the classes if directly evaluated on the gold data-set, suggesting that it was trained to fit the data patterns significantly different from those present in the collected data-set. However, pre-training on the silver data serves as a good starting point for further model fine-tuning. The obtained scores in general imply that there are many differences between the gold and silver data, including their vocabularies, answer distributions, and level of question detail.

7.3 Comparison of gold and silver data

To provide a meaningful analysis of the sources of difference between the gold and silver datasets,

Train data	'Yes'	'No'	Macro F1
Silver	0.00	0.00	0.00
Gold	0.27	0.77	0.35
Silver+gold	0.60	0.82	**0.47**

Table 6: Comparative performance of the SAN model trained on different combinations of silver and gold data, and evaluated on the test subset of gold data. Note that the gold annotations did not contain *'Not in report'* and *'Maybe'* options.

we grouped the gold questions semantically by using the CheXpert vocabulary for the 13 labels used for the construction of the silver dataset. The gold questions that are unable to be grouped via CheXpert were mapped manually using expert clinical knowledge. We systematically compared the gold and silver dialogs on the same 100 chest X-rays and noted the following differences.

- **Frequency of semantically equivalent questions**. Just under half of the gold question types were semantically covered by the questions in the silver dataset.
- **Granularity of questions**. We observed that the silver dataset tends to ask highly granular questions about specific findings (e.g. "consolidation") as expected. The radiology experts, however, asked a range of low (e.g. "Are there any bone abnormalities?), medium (e.g. "Are the lungs clear?") and high (e.g. "Is there evidence of pneumonia?") granularity questions. The gold dialogs tend to start with broader (low granularity) questions and narrow the differential diagnosis down as the dialogs progress.
- **Question sense.** The radiologists also asked questions in the form of whether some structure is "normal" (e.g. *"Is the soft tissue normal?"*). Whereas, the silver questions only asked whether an abnormality is present. Since chest X-rays are screening exams where a good proportion of the images may be "normal", having more questions asking whether different anatomies are normal would, therefore, yield more *'Yes'* answers.
- **Answer distributions** The answer distributions of the gold and silver data differ greatly. Specifically, while the gold data was composed heavily of *'Yes'* or *'No'* answers, the silver comprised mostly of *'Not in report'*.

8 Discussion

Our main finding is that the introduced task of visual dialog in radiology presents a lot of challenges from the machine learning perspective, including a skewed distribution of classes and a required ability to reason over both visual and textual input data. The best of our baseline models achieved 0.34 macro-averaged F1-score, indicating on a significant scope for potential improvements. Our comparison of gold and silver standard data shows some trends are in line with medical doctors' strategies in medical history taking, starting with broader, general questions and then narrowing the scope of their questions to more specific findings (Talbot et al.; Campillos-Llanos et al., 2020).

Despite the difficulty and the practical usefulness of the task, it is important to list the limitations of our study. The questions were limited to presence of 13 abnormalities extracted by CheXpert and the answers were limited to 4 options. The studies used in this work (from MIMIC-CXR) originate from a single tertiary hospital in the United States. Moreover, they correspond to a specific group of patients, namely those admitted to the Emergency Department (ED) from 2012 to 2014. Therefore, the data and hence the model reflect multiple real-world biases. It should also be noted that chest X-rays are mostly used for screening than diagnostic purposes. A radiology image is only one of the many data points (e.g. labs, demographics, medications) used while making a diagnosis. Therefore, although predicting presence of abnormalities (e.g. pneumonia) based on brief knowledge of the patient's medical history and the chest X-ray might be a good exercise and a promising first step in evaluating machine learning models, it is clinically limited.

There are plenty of directions for future work that we intend to pursue. To make the synthetic data more realistic and expressive, both questions and answers should be diversified with the help of clinicians' expertise and external knowledge bases such as UMLS(Bodenreider, 2004). We plan to enrich the data with more question types, addressing, for example, the location or the size of a given lung abnormality. We plan to collect more real life dialog between radiologists and augment the two datasets to get a richer set of more expressive dia-

log. We anticipate that bridging the gap between the silver- and the gold-standard data in terms of natural language formulations would significantly reduce the difference in model performance for the two setups.

Another direction is to develop a strategy to manage the uncertain labels such as *'Maybe'* and *'Not in report'* to make the dataset more balanced.

9 Conclusion

We explored the task of Visual Dialog for radiology using chest X-rays and released the first publicly available silver- and gold-standard datasets for this task. Having conducted a set of rigorous experiments with state-of-the-art machine learning models used for the combination of visual and language reasoning, we demonstrated the complexity of the task and outlined the promising directions for further research.

Acknowledgments

We would like to thank Mousumi Roy for her help in this project. We are also thankful to Mehdi Moradi for helpful discussions.

References

AB Abacha, SA Hasan, VV Datla, J Liu, D Demner-Fushman, and H Müller. 2019. 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.

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

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4971–4980.

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

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6077–6086.

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

Mythreyi Bhargavan, Adam H Kaye, Howard P Forman, and Jonathan H Sunshine. 2009. Workload of radiologists in united states in 2006–2007 and trends since 1991–1992. Radiology, 252(2):458–467.

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

Leonardo Campillos-Llanos, Catherine Thomas, Éric Bilinski, Pierre Zweigenbaum, and Sophie Rosset. 2020. Designing a virtual patient dialogue system based on terminology-rich resources: Challenges and evaluation. Natural Language Engineering, 26(2):183–220.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186.

Mark J Halsted and Craig M Froehle. 2008. Design, implementation, and assessment of a radiology workflow management system. American Journal of Roentgenology, 191(2):321–327.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778.

Drew A Hudson and Christopher D Manning. 2019. 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.

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. 2018. Overview of imageclef 2018: Challenges, datasets and evaluation. In International Conference of the Cross-Language Evaluation Forum for European Languages, pages 309–334. Springer.

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

Baoyu Jing, Pengtao Xie, and Eric Xing. 2018. On the automatic generation of medical imaging reports. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2577–2586.

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

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. Scientific data, 3:160035.

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

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

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

Mark D Mangano, Arifeen Rahman, Garry Choy, Dushyant V Sahani, Giles W Boland, and Andrew J Gunn. 2014. Radiologists' role in the communication of imaging examination results to patients: perceptions and preferences of patients. American Journal of Roentgenology, 203(5):1034–1039.

Yulei Niu, Hanwang Zhang, Manli Zhang, Jianhong Zhang, Zhiwu Lu, and Ji-Rong Wen. 2019. Recursive visual attention in visual dialog. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6679–6688.

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

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 1532–1543.

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

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1586–1596.

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. 2000. The earth mover's distance as a metric for image retrieval. International journal of computer vision, 40(2):99–121.

Bettina Siewert, Olga R Brook, Mary Hochman, and Ronald L Eisenberg. 2016. Impact of communication errors in radiology on patient care, customer satisfaction, and work-flow efficiency. American Journal of Roentgenology, 206(3):573–579.

Thomas B Talbot, Kenji Sagae, Bruce John, and Albert A Rizzo. Designing useful virtual standardized patient encounters.

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

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers. 2018. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 9049–9058.

Han Xiao. 2018. bert-as-service. https://github.com/hanxiao/bert-as-service.

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

Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Yin and yang: Balancing and answering binary visual questions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5014–5022.