Medical Image Captioning & VQA using Multitasking for Chest X-Rays

Disha Agarwal
dagarw24@asu.edu
Arizona State University
Tempe, Arizona, USA

Amey Bhilegaonkar
abhilega@asu.edu
Arizona State University
Tempe, Arizona, USA

Ninad Nale
nnale@asu.edu
Arizona State University
Tempe, Arizona, USA

Janaki Venkata Ramachandra Sai Nayani
jnayani@asu.edu
Arizona State University
Tempe, Arizona, USA

ABSTRACT

The interpretation of medical X-Ray images plays a crucial role in clinical decision-making. Machine learning-based approaches have shown great potential in assisting medical professionals in this task. In this project, we propose deep learning models and their comparison for both caption generation and Visual Question Answering (VQA) for medical X-Ray images. We aim to convert the task of captioning to prompt-based input. We will explore SOTA models such as Visual BERT for this task, and simple encoder-decoder models in addition to existing methods. The goal is to improve the interpretability and accuracy of the model, making it more useful for medical practitioners. The proposed approach is expected to address the challenges associated with detecting objects and identifying relationships between different objects in medical X-Ray images. The research in this domain has made a significant contribution to the field of medical image analysis, with potential applications in diagnostic and decision-making systems in clinical settings.

CCS CONCEPTS

• Computing Methodologies → Artificial Intelligence, Machine Learning.

KEYWORDS

Visual Question Answering (VQA), feature extraction, IUC Dataset, Convolutional Neural Networks (CNN), Attention mechanism, Semantic tags/labels.

1 INTRODUCTION

Image captioning and VQA both involve techniques from both computer vision and natural language processing. Captioning involves providing a concise summary of an image by identifying its important components, features, and relationships between them. Achieving this task requires the algorithm to generate grammatically correct text that captures the syntax and semantics of the summary. Deep learning is a powerful tool for tackling these challenges due to its layered structure and diverse architectures. Image captioning is essentially the combination of computer vision and natural language processing to recognize image components and generate human-readable sentences. While image captioning is an end-to-end process that encodes images into pixel sequences using an encoder and then decodes them into descriptive sentences using a decoder, it poses numerous challenges due to the complexity of mimicking the human brain's ability to understand and describe images. In the following section, we will explore these challenges in detail.

Visual Question Answering (VQA) involves answering a question based on the image input. Both modalities are necessary for a comprehensive solution to the VQA problem [6]. Existing VQA models provide single-word answers, but there is a need for a model that can provide a sequence of word answers, especially for medical images that require longer explanations [6]. In this project, we aim to explore the effectiveness of two popular convolutional neural networks (CNN) models, densenet121 and Xception, for image feature extraction. We will then use simple encoder-decoder models and VisualBERT which is a multimodal SOTA model.

This report is organized as follows. The related work section provides a review of the literature study and relative content that we found related to medical image captioning, VQA, and multitasking approaches. The dataset description section describes the chest X-ray dataset used in our study and the data pre-processing steps. The methodology section provides a detailed explanation of the approach used for medical image captioning and VQA tasks. In the results section, the findings of our study/experiments are presented, including evaluation metrics for the models. Finally, the conclusion section

summarizes the key findings of our study, limitations, and future research directions.

2 RELATED WORK

Several studies have been conducted on medical image captioning and VQA separately, but only a few have explored the multitasking approach to address both tasks simultaneously. In this section, we review the related works on medical image captioning and VQA, as well as the studies that have explored multitasking for these tasks in the context of chest X-rays.

2.1 Captioning & VQA

All the literature we studied has focused on biomedical image captioning as well as VQA using CNN to produce image features. The attention-based LSTM framework is employed in most of the articles to generate the output caption/answer. A SoftMax layer or completely linked layers on top of the LSTMs are used in certain works to solve the problem as a multilabel retrieval system, as in [8] and [3], respectively. Papers [4], [8] employ two LSTM layers, the first of which is referred to as Sentence LSTM (create a topic for a sentence), and the second of which is referred to as Word LSTM (provide the final output). To improve accuracy, [6] presents a combined approach of semantic tags or labels with other input characteristics. Attention is paid to both image and semantic tags and both are used as input to the LSTM layer. Our major focus is on the studies which used IUC dataset [2], [7] and ImageCLEF [6], to train and test the models. Multiple picture inputs are also used, such as lateral and anterior x-ray image characteristics in one article, and image features of a normal image and a patient image in another article.

2.2 Encoder-Decoder Framework

Encoder-Decoder Frameworks have been extensively studied and are a must to have in the fields of image captioning and VQA. In this framework, a Convolutional Neural Network (CNN) is used as an Encoder to extract image features, and a Recurrent Neural Network (RNN), such as Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU), is used as a Decoder to generate captions or answers. Attention mechanisms have also been added to improve the performance of these frameworks.

Several studies have used the Encoder-Decoder framework with attention mechanisms for image captioning and VQA. For example, [1], [2], [3], [4], [7], and [8] all use a CNN as the Encoder and an attention-based LSTM framework as the Decoder. Some studies have also used additional layers, such as a softmax layer or completely connected layers, on top of the LSTM to solve the problem as a multilabel retrieval system. [3] and [8] have used Faster-RCNN to generate the most relevant lesion or Region of Interest (ROI) for the output to have even more relevant image features. [7] uses a hierarchical RNN, LSTM with co-attention, and a transformer for VQA.

Moreover, some studies have employed multiple image inputs for captioning and VQA. For example, [2] uses lateral and anterior x-ray image characteristics as input, and [7] generates the final feature vector by comparing the image features of a normal image and a patient image. Some studies have also combined semantic tags or labels with other input features to improve accuracy. [1], [2], [3], [4], and [5] have all combined semantic tags or labels with image features and used both as input to the LSTM layer.

2.3 Evaluation Metrics

Evaluating the performance of image captioning and VQA models is a challenging task due to the subjective nature of language understanding and the lack of a standard for evaluation.

Various metrics are used to evaluate the performance of the models in biomedical picture captioning and VQA. The most commonly used metric is the BLEU score, which measures the similarity between the predicted caption and the reference caption in terms of n-grams (n-gram is a sequence of n words). Most of the articles, including [1], [2], [3], [4], and [7], use BLEU score as the primary evaluation metric. In addition to BLEU, other metrics used in the literature include ROUGE (Recall-Oriented Understudy for Gisting Evaluation), METEOR (Metric for Evaluation of Translation with Explicit Ordering), and CIDEr (Consensus-based Image Description Evaluation). There are some limitations of metrics such as BLEU, CIDEr. To understand better evaluation techniques, we studied Word Based Semantic Similarity. In [6], Word-Based Semantic Similarity (WBSS) is used as one of the evaluation metrics. WBSS measures the semantic similarity between the predicted and reference captions at the word level. The WBSS is calculated by computing the average of the cosine similarities between the word vectors of the predicted and reference captions. The experimental results in [6] show that the WBSS metric correlates well with human judgments of caption quality.

3 DATA

3.1 Dataset Description

In this paper, we are using 2 datasets which are predominantly for Image captioning and Visual Question and Answering. One is Indiana University chest X-rays data which is used by most research papers for Image captioning, the other is the Visual Question and Answer dataset for the VQA task.

3.1.1 Indiana University Chest X-ray: A dataset containing X-rays and medical reports given by the medical professionals associated with those X-rays. The reports contain

observations/ findings and conclusions based on observation which is the impression of the X-ray. A particular report may be mapped to more than one X-ray image. There are different orientations of the x-rays in the dataset like Frontal, lateral, etc. The dataset contains 7470 X-rays mapped with 3995 reports. We are using this dataset for the X-ray captioning task in our paper.

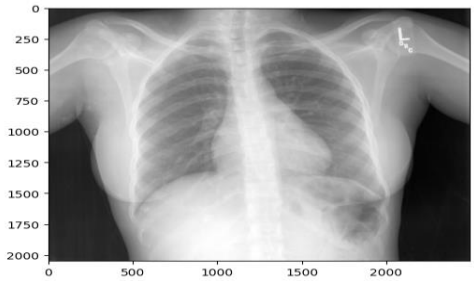

Figure 1: Example X-ray from Indiana University Chest X-ray and their findings.

Below is the information in the dataset that is associated with the record displayed in Figure 1.

- Problem: normal
- findings: The cardiac silhouette and mediastinum size are within normal limits. There is no pulmonary edema. There is no focal consolidation. There are no XXXX of a pleural effusion. There is no evidence of pneumothorax.
- impression: Normal chest x-XXXX.
- caption: Normal chest x-XXXX. The cardiac silhouette and mediastinum size are within normal limits. There is no pulmonary edema. There is no focal consolidation. There are no XXXX of a pleural effusion. There is no evidence of pneumothorax.

3.1.2 Visual Question and Answering Dataset: A question-and-answer image dataset, where Q/A pairs are mapped with an X-ray. This dataset is taken from Open-Source Framework (OSF) which is created manually where professionals are asked naturally occurring questions along with answers given by them. This dataset contains images of various organs including the chest, head, and abdomen. This dataset contains a total of 2248 radiology images with corresponding questions and answers. The answers include yes/no including a custom answer given by a professional.

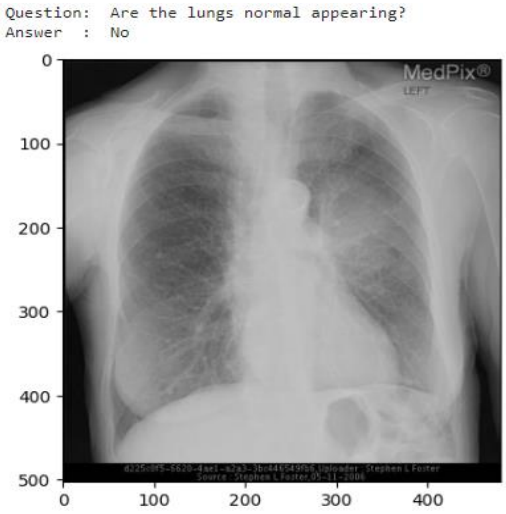

Figure 2: Example X-ray from VQA Dataset with Question and Answers.

3.2 Data Preprocessing

3.2.1 Data Cleaning: As we are using 2 different datasets, we need to obtain consistency across both, we are dropping all the other orientations of X-rays except for frontal X-rays. In the Visual Question and Answer data set, we are dropping all other organs except the chest X-rays. All the X-rays present in the VQA Dataset are frontal images. In the Indiana Chest X-ray dataset, we are keeping dropping all the columns and are keeping only findings and impressions. In the Virtual Question and Answering dataset, we are not dropping any columns.

3.2.2 Data Preparation: We are checking the various distributions of the dataset to get a better understanding of the data. We are also filtering the Visual Q & A with the chest to take only chest X-rays. We removed the columns Q_REPHRASE, and Q_FRAMED and added the questions a new entry for those images to have more questions for each image. In Indiana Dataset, we are combing both findings and report columns into the caption column and treating it as an answer for the VQA task. For the questions in the VQA, we are creating a prompt column and filling it with 2 questions: “Write a long caption of the given x-ray image” and “Write a short caption of the given x-ray image”. We are treating the caption as a long caption when findings are present and if findings are absent, we are treating it as a short caption. Figure 3 Shows the distribution of various how many X-rays that each report is mapped which is similar to mentioned in [9].

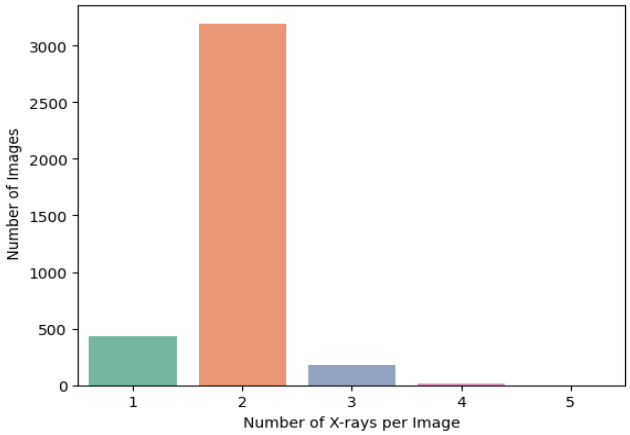

Figure 3: Count plot of the number of reports associated with each report [9].

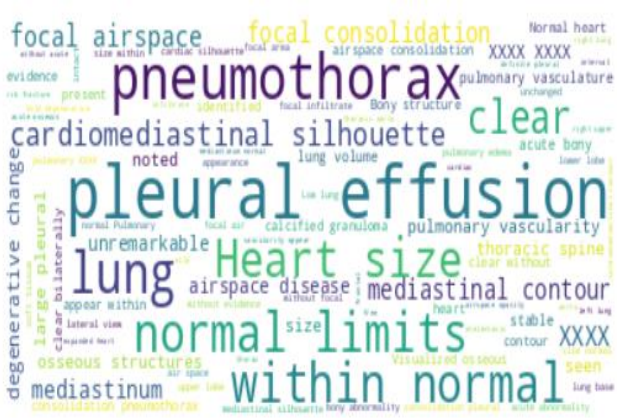

Figure 4: Word Cloud Indicating the most occurred terms in findings.

4 METHODOLOGY

To solve the concerned problem of generating Image captions and VQA tasks we use convolutional neural networks for encoding image features and LSTM to encode our prompt/question, for generating the output we use a decoder model. These are the two main components in the system, namely the encoder and the decoder which are neural networks trained in a supervised fashion. We trained different encoder architectures on the data set and selected the best performer according to the metrics used. A Decoder follows LSTM architecture and is trained with the features generated by the encoder and the image captions to generate image captions. Below we have detailed the components we have used in our system.

4.1 Feature Extraction

4.1.1 Image Feature Extraction: To ensure proper image feature extraction we are using the DenseNet121 and Xception architecture, it is imperative to resize the input images to specific dimensions. Specifically, images must be 224x224 pixels for DenseNet and 299x299 pixels for Xception. As for text descriptions, it is essential to preprocess them by removing punctuation marks and creating a dictionary of words. Subsequently, captions are converted into vectors using the dictionary indices to facilitate image comprehension. Thank you for your attention to this matter.

Xception Model: Xception is a CNN model pre-trained on ImageNet. It was developed by Google and is an Extreme version of Inception. The key architectural feature of Xception is depth wise Separable Convolutions. depth wise Separable Convolutions work with both spatial dimension and depth dimension.

DenseNet Model: DenseNet model is a CNN model, which is pre-trained on chexpert - which is a Chest X-ray dataset. It was developed by researchers at Facebook and Cornell University.

4.1.2 Text Pre-processing: First, the captions are cleaned by removing punctuation, and then a word dictionary is created. The dictionary indices are then used to convert the captions to vectors. We are attempting to sanitize the text dataset more accurately.

4.2 Simple Encoder-Decoder Model

We are using 3 LSTM layers to encode the prompt/question, for which we are - pre-processing text and forming a token vector using TFIDVectorizer from sklearn. We are generating the de-coded output using 3 LSTM decoders.

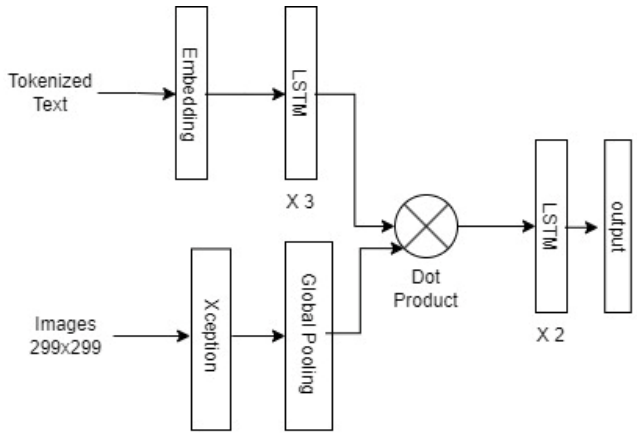

Figure 5: The simple encoder-decoder model.

The task of decoder is to generate image captions using the input of encoded image feature vectors. For decoder we are using Recurrent Neural Network specifically LSTM. LSTM stands for Long Short-Term Memory as they are capable of long-term dependencies. Also, LSTM has feedback connections which allow the processing of the entire sequence of data. The

Central role in the LSTM architecture is held by a memory cell also known as 'cell state'. Cell State allows the addition and removal of information and is controlled by Gates. The decoder consists of the image and language model. The image model has a dense 128-unit layer with Relu activation, and the LSTM has hidden layers of 128 and 512 units. The activation of the LSTM is softmax. The model is trained with RMSProp as the optimizer for a batch size of 256 for 50 epochs.

4.3 Visual BERT Model:

Visual BERT (VisBERT) is a multi-modal transformer architecture that combines the power of BERT with visual information for tasks that involve both language and images. The architecture of Visual BERT is similar to that of BERT, with the addition of a visual encoder that processes images. The Visual BERT model consists of three main components:

1. A text encoder: The text encoder is based on the BERT architecture and processes textual input to produce contextualized word embeddings. BERT uses a bidirectional transformer encoder to generate these embeddings, which are then used to produce representations for the entire sequence.
2. A visual encoder: The visual encoder is based on a ResNet architecture that processes image input to produce visual embeddings. The ResNet is pre-trained on the ImageNet dataset and fine-tuned on the target task. In our case, this image features from the CNN models will act as input to this section.
3. A multi-modal fusion layer: The fusion layer combines the text and visual embeddings to generate joint representations that capture the information from both modalities. The fusion layer includes two attention mechanisms: cross-modal attention, which allows the model to attend to both text and visual information when processing each modality, and self-attention, which enables the model to attend to different parts of the input sequence.

4.4 Evaluation

4.4.1 ROGUE: The ROUGE metric is a tool used to evaluate the performance of machine learning models designed to generate text. It stands for "Recall-Oriented Understudy for Gisting Evaluation," and it measures how well a model can automatically produce summaries or translations that are similar to human-generated summaries or translations.

ROUGE works by comparing the generated text to a set of reference texts, typically created by human experts. It then calculates the overlap between the generated text and the reference text, in terms of the number of shared words, sequences of words, and other linguistic units.

The quality of generated summaries, translations, and other types of text output is frequently evaluated in the fields of natural language processing and machine translation research using the ROUGE metric. By comparing many models and choosing the one that yields the highest ROUGE scores, it can also be used to optimize models' performance.

4.4.2 BLEU Score: BLEU Rating the Bilingual Evaluation Understudy Score, or BLEU, is a statistic for comparing a sentence that was generated to a sentence that was used as a reference. A perfect match is represented by a score of 1, whereas a perfect mismatch is represented by a score of 0. The score was developed in order to evaluate the predictions made by autonomous machine translation systems. Although it is not perfect, it has five very good benefits:

1. Calculation is quick and affordable.
2. It is simple to comprehend.
3. It is linguistically unrestricted.
4. It has a strong correlation with human evaluation.
5. It has received a lot of adoptions.

$$BLEU = BP \exp(\Sigma_{n=1}^{N} w_n log p_n)$$

Where, BP is Brevity Penalty
N = the total number of n-grams, such as unigrams, bigrams, 3-grams, and 4-grams
w_n = the weight of each modified precision
p_n = the modified precision

5 EXPERIMENTS

Baselines: For producing baselines we will train the simple encoder-decoder model on single task i.e. VQA and Captioning separately and test it on the task it is trained on. We will also finetune Visualbert on these tasks and test the performance individually (depending on time and resources).

Multitasking: For this section we will train the simple encoder-decoder on both the tasks and test on both the tasks. We will also finetune VisualBERT on these tasks and test the performance (depending on time and resources).

If time permits, we will also try to modify our simple encoder decoder model to an attention-based model.

6 RESULTS

Over the next week, we will complete the experiments on all our proposed models and will finalize the results section. [TBD]

7 CONCLUSION

Once we have completed implementation of all our models, we will evaluate them and based on all the results we will draw conclusions about our findings. These will be presented in the final report. [TBD]

ACKNOWLEDGMENTS

We would like to express our gratitude to Dr. Hannah Kerner and the teaching assistant Mirali Purohit for providing us with this wonderful opportunity to work on the paper. Their invaluable feedback has significantly contributed to the completion of this work. We are grateful to all the authors whose work and literature we have followed to work on this project.

REFERENCES

[1] Xu, J., Huang, J., Wang, C., Zhang, C., & Wang, J. (2021). Automatic ultrasound image report generation with adaptive multimodal attention mechanism. Artificial Intelligence in Medicine, 114, 102051. DOI: 10.1016/j.artmed.2020.102051.

[2] Yang, S., Niu, J., Wu, J., Liu, X. (2020). Automatic Medical Image Report Generation with Multi-view and Multi-modal Attention Mechanism. In: Qiu, M. (eds) Algorithms and Architectures for Parallel Processing. ICA3PP 2020. Lecture Notes in Computer Science(), vol 12454. Springer, Cham. https://doi.org/10.1007/978-3-030-60248-2_48

[3] Zeng, X., Wen, L., Xu, Y., & Ji, C. (2020). Generating diagnostic report for medical image by high-middle-level visual information incorporation on double deep learning models. Computer Methods and Programs in Biomedicine, 197, 105700. https://doi.org/10.1016/j.cmpb.2020.105700

[4] Yin, C., Zhou, Y., Zhang, X., Zhang, Y., Xu, D., & Wang, Y. (2019). Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network. In 2019 IEEE International Conference on Data Mining (ICDM) (pp. 728-737). DOI: 10.1109/ICDM.2019.00086

[5] M. Gu, X. Huang and Y. Fang, "Automatic Generation of Pulmonary Radiology Reports with Semantic Tags," 2019 IEEE 11th International Conference on Advanced Infocomm Technology (ICAIT), Jinan, China, 2019, pp. 162-167, DOI: 10.1109/ICAIT.2019.8935910.

[6] R. Ambati and C. Reddy Dudyala, "A Sequence-to-Sequence Model Approach for ImageCLEF 2018 Medical Domain Visual Question Answering," 2018 15th IEEE India Council International Conference (INDICON), Coimbatore, India, 2018, pp. 1-6, DOI: 10.1109/INDICON45594.2018.8987108.

[7] Hyeryun Park, Sanghoon Park, and Jeongyeon Kim. 2021. Medical Image Captioning Model to Convey More Details: Methodological Comparison of Feature Difference Generation. IEEE Access 9 (2021), 150560–150568. DOI: https://doi.org/10.1109/ACCESS.2021.3124564.

[8] Anderson, P. et al. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR '18). IEEE, Salt Lake City, UT, USA, 6077-6086. DOI: https://doi.org/10.1109/CVPR.2018.00636.

[9] Rohan Soni. (2019, November 25). Indiana University Chest X-Rays Automated Report Generation. Medium. https://rohansoni-jssaten2019.medium.com/indiana-university-chest-x-rays-automated-report-generation-38f928e6bfc2.