Medical Image Captioning Diagnostic Report Generation

Srijan Sabbineni[a], Nagendra Parvathaneni[a], Nandini Kodali[a], Sankar Jai Pavan Uttaravalli[a]

[a]*Department of Computer Science and Engineering, SRM University, Amaravati, 522240, Andhra Pradesh, India*

Abstract

This report provides an overview of how Artificial Intelligence is used to generate descriptions and diagnostic reports from medical images such as chest X-rays. It reviews the key techniques used in Natural Image Captioning, Medical Image Captioning, and Diagnostic Captioning, including CNN-based encoders, attention mechanisms, and Transformer models. The report also summarizes major medical datasets like IU X-ray, PadChest, and MIMIC-CXR, and explains how these models are evaluated using both text-based and clinical metrics. Finally, it highlights the main challenges—such as limited data, hallucinations, and clinical accuracy—and outlines future directions for improving automated medical reporting systems.

Keywords: Image Captioning, Medical Image Captioning, Diagnostic Captioning, Radiology Reports, Encoder–Decoder, Attention Mechanisms, Transformers, Chest X-ray Datasets, Clinical Evaluation

1. Introduction

Medical Image Captioning is an emerging research area that combines computer vision and natural language processing to automatically generate descriptive text for medical images such as chest X-rays. Traditional image captioning methods from natural image datasets have evolved into more advanced approaches for medical applications, focusing on clinical accuracy and reliability.

Diagnostic Captioning further extends this by producing structured radiology-style reports that support clinical decision-making. With growing imaging volumes and limited radiologist availability, automated captioning systems offer potential benefits such as reducing workload, improving consistency, and assisting healthcare settings with limited expertise.

Email addresses: venkatasatyasrijan_sabbineni@srmap.edu.in (Srijan Sabbineni), himanagendra_parvathaneni@srmap.edu.in (Nagendra Parvathaneni), nandini_kodali@srmap.edu.in (Nandini Kodali), sankarjaipavan_uttaravalli@srmap.edu.in (Sankar Jai Pavan Uttaravalli)

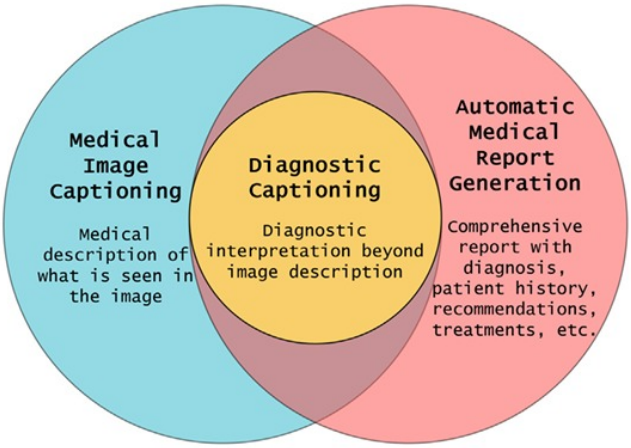

Figure 1: Visual Representation of the definition of Medical Image Captioning(MIC),Diagnostic Captioning(DC), and automatic medical report generation.

This review summarizes key methods, datasets, evaluation metrics, and challenges involved in developing accurate and clinically meaningful captioning systems.

2. Objective

• To study the evolution of image captioning from Natural Image Captioning (NIC) to Medical Image Captioning (MIC) and Diagnostic Captioning (DC).

• To analyze various deep learning approaches used for generating medical image descriptions and radiology-style diagnostic reports.

• To review major large-scale medical datasets employed for training medical captioning models.

• To examine evaluation metrics that assess both linguistic quality and clinical correctness.

• To summarize key challenges, limitations, and potential improvements for the real-world deployment of MIC and DC systems.

3. Significance

Medical imaging volumes are increasing rapidly, leading to heavy workloads for radiologists and potential delays in generating diagnostic reports. Automated captioning systems can assist by producing faster and more consistent interpretations of medical images, especially in settings with limited clinical resources.

Radiology reports play a central role in clinical decision-making, as they provide a structured description of findings, comparisons, and diagnostic impressions. Early AI-generated draft reports can support clinicians by highlighting key observations, reducing reporting time, and improving workflow efficiency.

Understanding existing captioning methods helps identify current limitations and guides the development of safer and more reliable diagnostic AI models. This study aims to enhance the accuracy, clarity, and clinical usefulness of automatically generated medical reports.

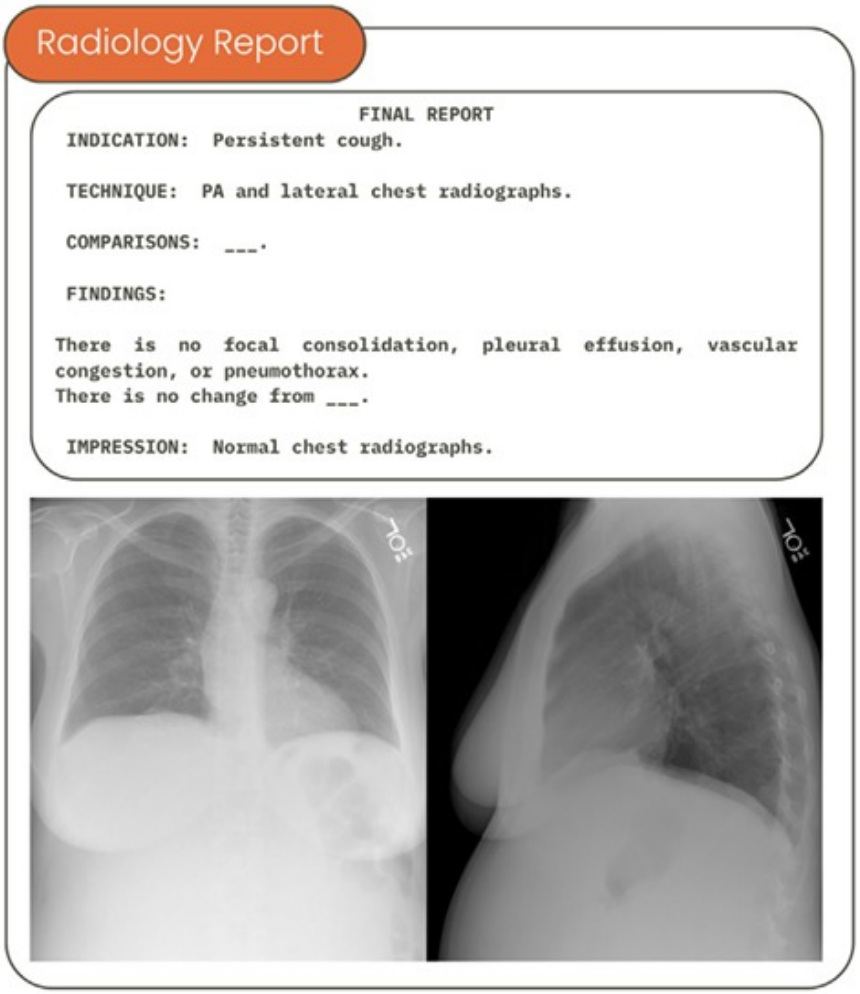

Figure 2: Example of a structured radiology report with corresponding chest X-ray images.

4. Research Method

The review followed a structured methodology to collect, screen, and analyze research on Natural Image Captioning (NIC), Medical Image Captioning (MIC), and Diagnostic Captioning (DC). Relevant studies published between 2002 and 2023 were gathered from major scholarly databases such as Web of Science, Scopus, Google Scholar, PubMed, and arXiv with most of them from the year 2018.

Research papers were selected by applying keywords such as "image captioning," "medical image captioning," "radiology report generation," "deep learning," "CNN," "LSTM," "attention," and "transformers." The shortlisted studies were then organized according to model architecture, datasets utilized, and evaluation metrics.

NIC-related works were reviewed to understand the core encoder–decoder frameworks, whereas MIC and DC papers were analyzed to explore their domain-specific modifications, challenges, and clinical needs.

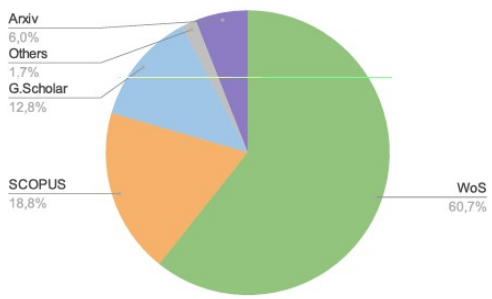

Figure 3: Database Distribution

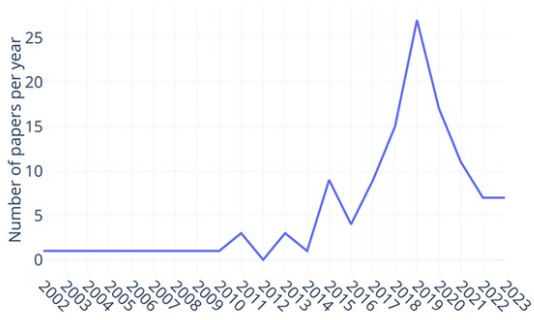

Figure 4: Year distribution

This systematic review enabled a comparative understanding of how captioning methods evolved from general vision–language models to clinically oriented diagnostic systems.

5. Datasets Used:

5.1. IU X-Ray Dataset:

- Contains around 8,000 chest X-rays paired with radiology reports. It is widely used for baseline captioning experiments.

5.2. PadChest Dataset:

- A large dataset of 160,868 chest X-rays with 109,931 Spanish/Valencian reports. About 27% of the labels were manually annotated by physicians, and the rest were generated using RNN-based automated labeling.

5.3. MIMIC-CXR Dataset:

- The largest publicly available chest X-ray dataset with 372,000 radiographs and 227,000 detailed reports. Includes structured labels such as CheXpert findings and supports clinical evaluation studies.

5.4. ROCO Dataset:

- It provides diverse multimodal image–text pairs that help train models to understand medical terminology and visual contexts.

5.5.

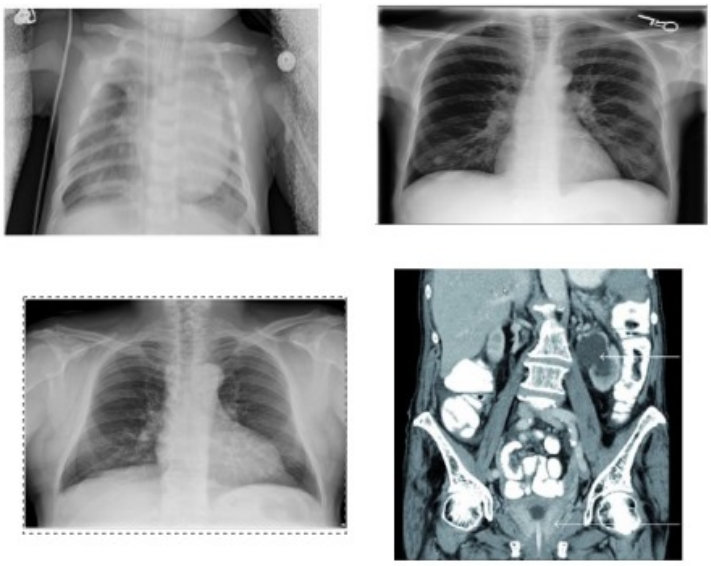

Figure 5: Samples of (a) The Indiana University Chest X-ray Collection (IU X-ray), (b) Pathology Detection in Chest Radiographs (PadChest dataset), (c) Medical Information Mart for Intensive Care - Chest X-ray (MIMIC-CXR), (d) Radiology Objects in COntext (ROCO)

6. Evaluation metrics:

Evaluation metrics measure how well the AI-generated medical captions match real radiology reports. Linguistic metrics such as BLEU, ROUGE, METEOR, and CIDEr compare the generated text with reference reports to check fluency and similarity. Clinical metrics like CheXpert and RadGraph focus on whether the medical findings and relations are correctly identified. These metrics together help assess both the language quality and the clinical accuracy of the model. A good captioning system should score well in both categories to ensure trustworthy diagnostic output.

7. Results:

The reviewed studies show that image captioning performance has steadily improved from traditional CNN–LSTM models to advanced Transformer-based architectures. Models using attention mechanisms and hierarchical decoding produce more coherent and clinically relevant descriptions compared to earlier approaches. Transformer models demonstrate superior linguistic accuracy on metrics such as BLEU, ROUGE, and CIDEr, while clinically focused metrics like CheXpert and RadGraph confirm better alignment with real diagnostic findings. Overall, recent approaches achieve stronger descriptive quality, improved disease detection consistency, and greater potential for generating structured radiology-style reports.