

Medical Image Captioning

Nagendra, Prof .Sivapuram Arunkumar
Department of CSE, SRM University AP

Abstract

Medical Image Captioning (MIC) and Diagnostic Captioning (DC) aims to describe medical images and generate **radiology-style reports automatically**. These systems integrate **deep learning, computer vision, and natural language processing** techniques to interpret medical images—particularly **chest X-rays**—and produce accurate descriptive or diagnostic text.

This review paper examines the progression from **Natural Image Captioning (NIC)** to **MIC** and **DC**, highlighting the differences in domain complexity and clinical requirements. It provides an overview of major benchmark datasets, including **IU X-Ray** and **MIMIC-CXR**, and analyzes current state-of-the-art models such as **encoder-decoder architectures, attention mechanisms, and Transformer-based frameworks**.

The paper also summarizes key challenges faced by MIC systems, including the **limited availability of annotated datasets**, and the need for **clinical correctness** and reliability. Additionally, it reviews widely used evaluation metrics, ranging from standard language metrics like **BLEU** and **CIDEr** to clinically oriented metrics such as **CheXpert** and **RadGraph**.

Objectives

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.

Significance

- Medical imaging volume is increasing rapidly, leading to high workload for radiologists and reporting delays.
- Automated captioning systems can enable faster and more consistent interpretation of medical images.
- Early AI-generated descriptions can support clinical decision-making, especially in resource-limited hospitals.
- Understanding existing methods helps identify gaps and guides the development of safer, more reliable diagnostic AI models.
- This study aims to improve the quality, accuracy, and clinical applicability of AI-generated medical reports.

Research Methodology

- Relevant studies from 2010–2023 were collected from **WoS, Scopus, Google Scholar, ArXiv**, and other related academic sources.
- Using defined **NIC, MIC, and DC keywords**, publications were systematically searched (Fig 2) and screened using **inclusion-exclusion criteria** (Fig 1)
- A total of 114 studies were selected, with most publications appearing in 2018, showing a peak in research activity.
- Each selected study was classified based on **model type, dataset, and evaluation metrics**.
- This process enabled a **structured comparison of methods, research progress, and emerging trends in medical image captioning**.

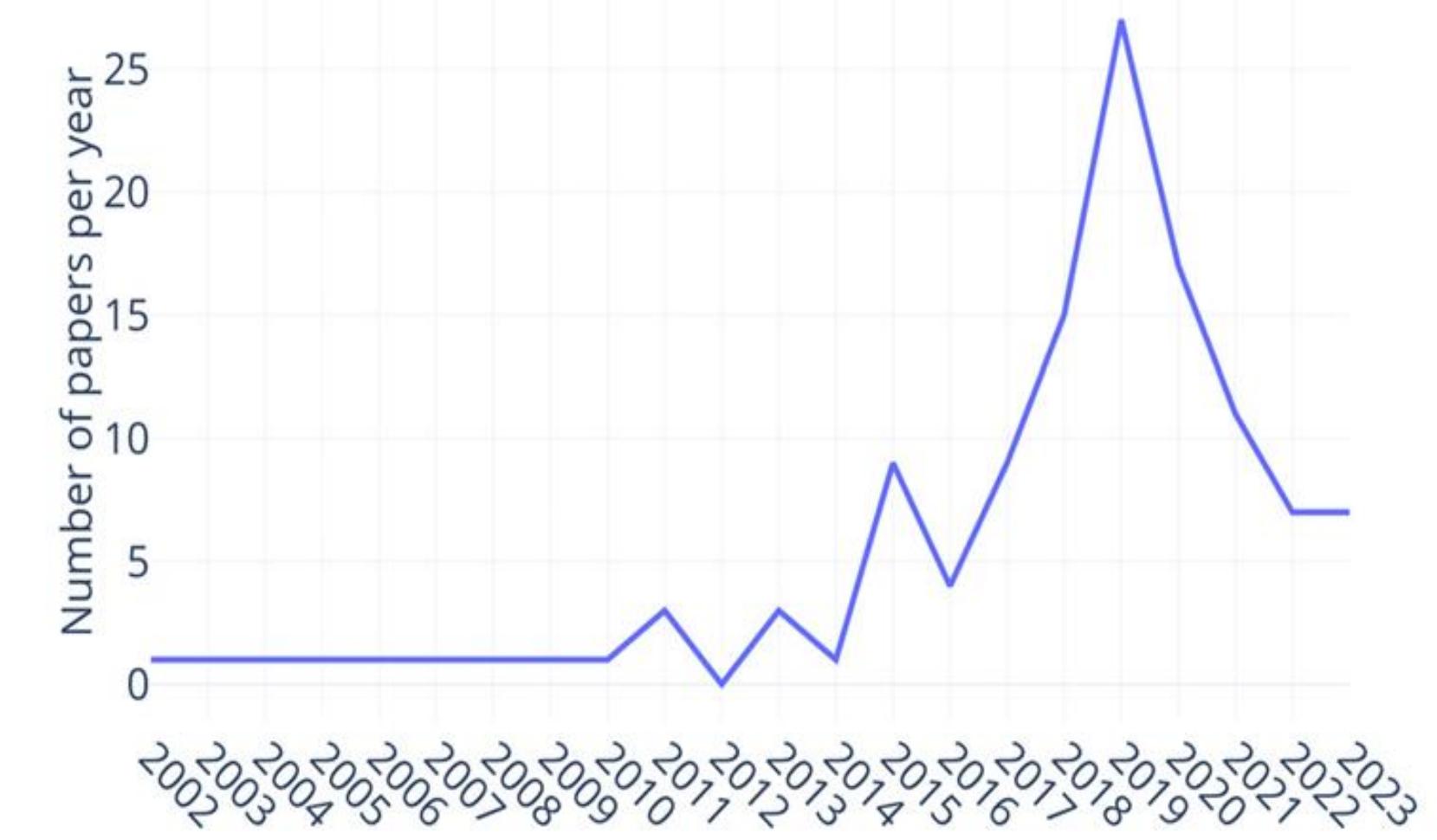


Figure1. Yearly Distribution

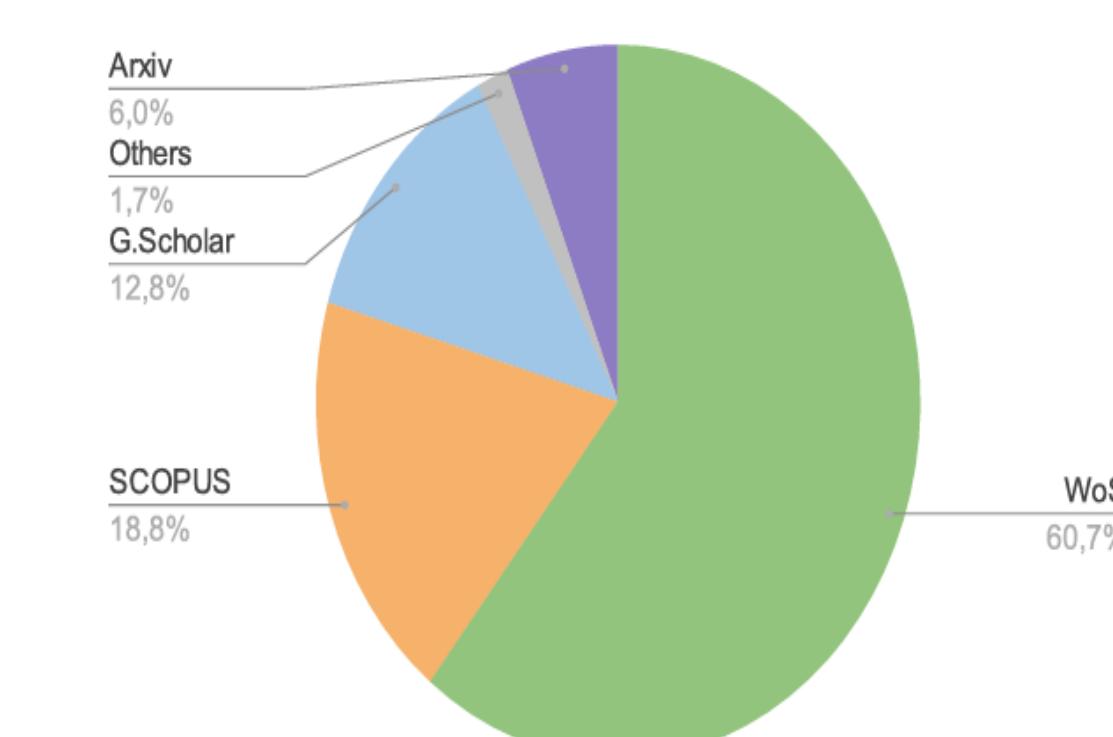


Figure2. Database Distribution

Methods and Materials

- Datasets Used:** The paper reviews major medical captioning datasets such as **IU X-Ray, MIMIC-CXR, PadChest, and ROCO**. These datasets provide **paired radiographs and reports** essential for training captioning models.
- Model Approaches:** The study highlights multiple architectures, including **CNN-RNN encoder-decoder models, attention mechanisms, hierarchical LSTMs** for generating long radiology reports, and **Transformer-based captioning models**. It also covers **dense captioning models** designed to detect and describe multiple abnormalities.
- Preprocessing:** Medical images undergo **normalization, cleaning, and resizing**. Reports are processed through **tokenization, label extraction, and vocabulary construction**.
- Training Pipeline:** Models are trained using **supervised learning with cross-entropy loss, along with reinforcement learning techniques such as CIDEr optimization** to improve clinical-text alignment.

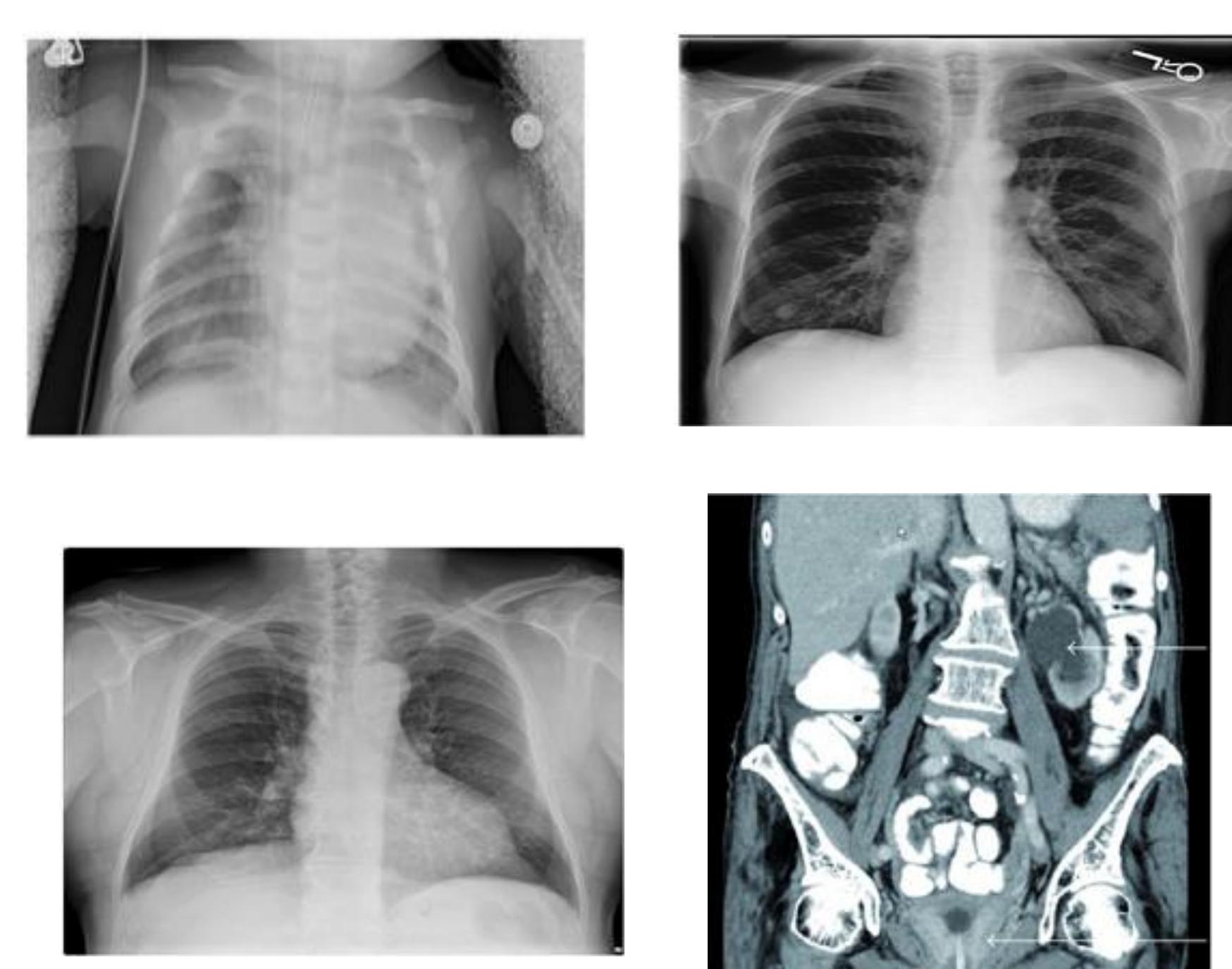


Figure3. Samples of (a) The Indiana University Chest X-ray Collection (IU X-ray),
(b) Pathology Detection in Chest Radiographs (PadChest dataset),
(c) Medical Information Mart for Intensive Care-Chest X-ray (MIMIC-CXR),
(d) Radiology Objects in COntext (ROCO)

Contribution(Diagnostic Captioning):

- Studied advanced **Diagnostic Captioning (DC)** models that generate full radiology reports, not just short captions.
- Analyzed **hierarchical decoders** designed for multi-sentence and **paragraph-level report generation**.
- Reviewed **region-based and dense captioning** approaches that focus on localized abnormalities within specific image regions.
- Studied **hybrid retrieval + generative models**, such as **KERP and HRGR-Agent**, which combine **template retrieval with neural generation** for more accurate reports.
- Examined **clinical reasoning components**, including how models generate **structured sections like Findings and Impression**.
- Compared **DC performance with MIC**, evaluating improvements in **diagnostic accuracy, clinical relevance, and report completeness**.

Results

- Transformer-based models demonstrated **higher caption accuracy** compared to **CNN-RNN** and **attention-only** architectures.
- Dense captioning models improved the ability to detect and describe **multiple abnormalities** across different regions of the image.
- Natural language evaluation scores (e.g., **BLEU**) increased as model complexity and representational power improved.
- Clinical metrics confirmed **more consistent and disease-specific outputs**.
- Retrieval-based systems generated **stable but less detailed reports**, while generative models produced **richer descriptions** but sometimes introduced **hallucinations**.
- Overall, Transformer models provided the **best balance of fluency, accuracy, and clinical correctness** in medical image captioning.

Reference

- Reale-Nosei, G., et al. (2024). From Vision to Text: A Review on Medical Image Captioning and Radiology Report Generation. *Medical Image Analysis*.
- Demner-Fushman, D., et al. (2016). IU Chest X-ray Dataset: A collection of X-rays with paired radiology reports.
- Johnson, A., et al. (2019). MIMIC-CXR: A large publicly available chest X-ray dataset with structured labels.
- Bustos, A., et al. (2020). PadChest Dataset: Large-scale labeled chest X-ray database for pathology detection.