Medical Diagnosis and Report Generation for Endoscopic Procedures

Praval Pattam

Potluri Theenesh

Pranav Sai Sarvepalli

Introduction

Doctors spend a significant portion of their workday on documentation, often dedicating 35-37% of their time to paperwork rather than direct patient care. This administrative burden not only contributes to physician burnout but also reduces the time available for meaningful patient interactions. AI-powered medical documentation tools, such as Vision-Language Models (VLMs), offer a transformative solution by automating report generation and enhancing diagnostic accuracy.

Recent advancements in multimodal learning have significantly reshaped healthcare AI, particularly with the rise of VLMs. These models integrate visual and textual data, enabling deep analysis and improved performance across various medical applications. State-of-the-art (SOTA) models such as CLIP, LLaVa, and Flamingo have been successfully adapted to the healthcare domain through extensive medical dataset training.

By leveraging a system built using these VLM, clinicians can:

- Pose queries about medical images to gain deeper insights.
- Generate contextually rich reports, streamlining medical documentation.
- Enhance diagnostic accuracy, reducing human error.
- Improve efficiency, allowing doctors to focus more on patient care.

This system not only reduces the time spent on report generation but also helps doctors read and interpret diagnoses more effectively, ensuring better patient outcomes and optimized workflows

Project Objective

This project aims to develop an **AI-powered automated system** for **endoscopic report generation** using VLMs. The system will:

- **Analyze endoscopic videos and images** to extract relevant medical insights.
- **Identify the patient's diagnosis** before generating a structured medical report.
- **Produce a detailed, specific endoscopic report** reflecting the nuances in the patient's condition.
- **Integrate a chat feature** to allow for **interactive queries**, enabling healthcare professionals to obtain additional insights.

By **combining textual and visual data**, this AI system ensures comprehensive documentation, enhances **clinical efficiency**, and improves **decision-making workflows**.

Dataset & Approach

- **Data Availability:** We utilize a dataset containing **2.5k folders with endoscopic images** and their **corresponding medical reports** which we received from GMC.
- **Processing Pipeline:**
  1. **Input Processing:** Endoscopic videos and images are fed into the system.
  2. **Feature Extraction:** The VLM identifies relevant **medical patterns** from visual data.
  3. **Diagnosis Identification:** The model first determines the **specific condition affecting the patient**.
  4. **Report Generation:** The system **translates extracted insights** into **structured medical documentation**.
  5. **Interactive Insights:** A **chat-based feature** enables **real-time clinician queries**, offering **additional analysis** of the findings.

Enhancing Medical Report Generation with Vision-Language Models (VLMs)

Expanded Overview of VLMs

Vision-Language Models (VLMs) are revolutionizing medical AI by integrating computer vision and natural language processing to analyze both visual and textual data. These models leverage deep learning techniques, particularly transformer-based architectures, to extract meaningful insights from medical images and generate structured reports. By fusing multimodal data, VLMs enhance diagnostic accuracy, streamline documentation, and improve clinical workflows.

Several state-of-the-art (SOTA) VLMs have been adapted for medical applications, including:

- **CAT-ViL DeiT** – Trained on **EndoVis 2017** and **2018**, specializing in robotic surgery analysis.
- **MedViLL** – A model designed for medical image-text understanding.
- **BiomedCLIP** – A specialized adaptation of CLIP for biomedical imaging.
- **LLaVa-Med** – A healthcare-focused version of LLaVa optimized for medical report generation.
- **Med-Flamingo** – A multimodal model tailored for medical image interpretation.

Availability of Public Datasets for our use case

One of the key advantages of developing AI-powered medical report generation systems is the availability of publicly accessible datasets, which eliminates the need for additional manual report collection. Several well-curated datasets provide extensive medical image-text pairs, enabling robust model training and validation. Some notable datasets include:

- **EndoVis** – A dataset focused on endoscopic vision tasks, including robotic surgery videos and annotated medical images.
- **ROCO** – A dataset featuring various medical imaging modalities with textual descriptions.
- **MedICaT** – A collection of medical images annotated with captions and structured reports.
- **VQA-Med** – A dataset designed for medical visual question answering tasks.

More about EndoVis Datasets (What we will probably use)**:**

Endoscopic Vision (EndoVis) 2017

The EndoVis 2017 dataset contains five robotic surgery videos with varying frame counts (8, 18, 14, and 39 frames) from the MICCAI Endoscopic Vision 2017 Challenge. It includes 472 QA pairs with bounding box annotations, carefully crafted to involve specific inquiries related to surgical procedures. Example questions include:

- *What is the state of prograsp forceps?*
- *Where is the large needle driver located?*

The inclusion of bounding box annotations enhances the dataset's utility for tasks such as object detection and answer localization.

EndoVis 2018

The EndoVis 2018 dataset expands on its predecessor, containing 14 robotic surgery videos (totaling 2,007 frames) from the MICCAI Endoscopic Vision 2018 Challenge. It includes 11,783 QA pairs regarding:

- Organs
- Surgical tools
- Organ-tool interactions

Future Prospects

Beyond initial **prognosis** and **report generation**, the system will evolve to:

- **Expand its dataset** to ensure greater **medical diversity** for improved diagnostic accuracy.
- **Validate its feasibility in clinical settings**, optimizing real-world applications.

References:

Vision Language Models in Medicine

https://arxiv.org/pdf/2503.01863

Vision-language models for medical report generation and visual question answering: a review

https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2024.1430984/full

Paper - https://arxiv.org/pdf/2403.02469

github link for the paper - https://github.com/lab-rasool/Awesome-Medical-VLMs-and-Datasets

Vision–Language Model for Visual Question Answering in Medical Imagery

https://www.mdpi.com/2306-5354/10/3/380

Large language models: a primer and gastroenterology applications

https://pmc.ncbi.nlm.nih.gov/articles/PMC10883116/

Supplementary material - shows the usecases -

https://pmc.ncbi.nlm.nih.gov/articles/PMC10883116/table/table1-17562848241227031/

USING GPT4V, A VISION-LANGUAGE LARGE LEARNING MODEL (LLM), TO GRADE THE BOSTON BOWEL PREPARATION SCORE

https://www.giejournal.org/article/S0016-5107(24)00929-5/abstract (ask sir for this paper)

Deep sight: enhancing periprocedural adverse event recording in endoscopy by structuring text documentation with privacy-preserving large language models

https://www.sciencedirect.com/science/article/pii/S2949708624001067

Revolutionizing gastrointestinal endoscopy: the emerging role of large language models

https://www.researchgate.net/publication/383546640_Revolutionizing_gastrointestinal_endoscopy_the_emerging_role_of_large_language_models

Large Language Models in Gastroenterology: Systematic Review

https://www.researchgate.net/publication/387273607_Large_Language_Models_in_Gastroenterology_Systematic_Review

The Potential Clinical Utility of the Customized Large Language Model in Gastroenterology: A Pilot Study

https://www.researchgate.net/publication/387371185_The_Potential_Clinical_Utility_of_the_Customized_Large_Language_Model_in_Gastroenterology_A_Pilot_Study