Medical Diagnosis and Automated Report Generation for Endoscopic Procedures

Praval Pattam,
Potluri Theenesh,
Pranav Sai Sarvepalli

In modern healthcare, doctors spend a substantial portion of their workday, estimated between 35–37%, on documentation and administrative tasks, which not only reduces the time available for direct patient interaction but also contributes significantly to doctor burnout and overall inefficiency in clinical workflows. The emergence of artificial intelligence (AI), particularly through Vision-Language Models (VLMs), offers a promising solution to this challenge by enabling automated analysis and report generation from medical imaging data, while simultaneously enhancing diagnostic accuracy and reducing human error. VLMs leverage state-of-the-art deep learning architectures that integrate visual and textual information, allowing them to analyze endoscopic images and videos comprehensively, extract clinically relevant features, and generate structured medical reports that capture nuances of the patient's condition. Several specialized VLMs have been successfully adapted to the healthcare domain, including CAT-ViL DeiT, designed for robotic surgery analysis, MedViLL for medical image-text understanding, BiomedCLIP for biomedical imaging, LLaVa-Med for healthcare-focused multimodal reasoning, and Med-Flamingo for image-based medical interpretation. The proposed system for endoscopic procedures utilizes these models to provide multiple functionalities: it can automatically identify specific patient diagnoses, generate detailed structured reports, and offer interactive capabilities through a chat-based interface, allowing doctors to query images and obtain additional insights in real time. The dataset for this project includes over 2,500 endoscopic image folders and associated medical reports obtained from GMC, supplemented by publicly available datasets such as EndoVis 2017 and 2018, which provide annotated surgical videos, bounding box labels, and QA pairs focused on organ-tool interactions and procedural steps. Additional datasets such as ROCO, MedICaT, and VQA-Med further enhance training and validation by providing diverse medical image-text pairs across multiple modalities. The processing pipeline involves several steps: input processing of endoscopic videos and images, feature extraction using VLMs to identify relevant medical patterns, diagnosis determination, report generation translating insights into structured documentation, and interactive insights for real-time doctor queries. This multimodal AI approach ensures comprehensive documentation, improves workflow efficiency, enhances diagnostic precision, and supports better patient outcomes. Looking forward, the system aims to expand its dataset diversity to cover a wider range of medical conditions, validate its performance in clinical settings to ensure practical applicability, and refine real-world deployment strategies, positioning AI-assisted endoscopic report generation as a key advancement in modern healthcare, reducing doctor workload and improving patient care quality at scale.