VisiHealth AI

Making Medical AI Human-Friendly

Supervisor:
Sir Abdul Rahman

Submitted by:
Junaid Mohi Ud Din (01-134222-071)
Hammad Ur Rehman (01-134222-059)

Department of Computer Science Bahria University, Islamabad

Chapter 1: Introduction

1.Introduction

1.1 Project Overview

VisiHealth AI is a multimodal Medical Visual Question Answering (Med-VQA) which aims to bring artificial intelligence and clinical trust closer to each other. The system combines computer vision and natural language processing features to not only provide answers to clinical questions using medical images but it also includes giving rationales based on medical knowledge. The system is able to generate verifiable explanations by decoding visual features of radiology images and interpreting textual clinical queries, thereby resolving the problem of black box nature of existing medical AI models.

The classic Med-VQA systems are usually aimed at giving straight forward answers without any evidence that the medical practitioners cannot rely on in such an emergency. To address them, this project makes use of the SLAKE dataset, which is compiled of radiology photographs, semantic segmentation masks, and structured medical knowledge graph. The VisiHealth AI uses a CNN that has been trained fresh to extract image features and localize regions of interest (ROI) as well as a fine-tuned BioBERT trained upon textual features.

The solution suggested combines these textual and visual inputs to deduce clinical responses and at the same time extracts pertinent information in a medical knowledge graph. The outcome of this is the production of a system giving a prediction and a chain of reasons, which focuses on interpretability and clinical relevance. Essentially, VisiHealth AI is a major advance in Explainable AI (XAI) in healthcare, and this will form a basis to AI-enforced medical diagnosis and education.

1.2 Problem Description

The issue of AI systems interpretability is mandible in the healthcare sector as much as it is the case in terms of accuracy. Although currently Med-VQA studies have had immense performance in tasks with data sets such as VQA-RAD or VQA-Med, these networks are primarily viewed as a black box. The yes without giving the answer, which poses an obstacle

to apply it in a medical practice where evidence is needed.

The fundamental issues that can be concluded upon are:

Lack of Explainability: The state of the art Med-VQA models give answers, but do not provide rationales or evidence to justify their answer.

Trust Factor: An AI system that only delivers the answer of Abnormality detected will never be trusted by medical professionals who do not indicate the particular image or medical fact, which resulted in such response.

Limitations of Data Sets: It is notable that lots of data sets do not have the semantic data or knowledge graph triplets to train models based on reasoning.

Lack of Context: Typical models do not consider any of the underlying medical expertise needed to relate visual symptoms with diagnostic terms.

VisiHealth AI can solve these problems by introducing a reasoning aware architecture. In comparison to ordinary classifiers, the system bases its responses on particular image regions (ROIs) and confirmed medical facts, and as a result of this, the decision making process of the AI is not only transparent, verifiable, but also clinically significant.

1.3 Project Objectives

The main purpose of the proposed project is to create and introduce a multimodal Med-VQA system enhancing trust in AI-supported healthcare because it provides both correct responses and readable answers. The following are the key objectives:

i. To create a hybrid Med-VQA architecture to combine both computer vision and natural language processing methods.

ii. To apply a ROI detector to identify and mark distinct organs or abnormalities in the medical images.

iii. To incorporate a Knowledge Graph retrieval module that retrieved the relevant medical fact with reference to the identified visual features and key questions.

iv. To build a generation engine rationale model that can combine retrieved facts in the form of KG and localized image areas into well-formed explanation.

v. To obtain grounded reasoning, to train and evaluate the model on the basis of the SLAKE dataset, one needs to use its semantic labels and knowledge triplets.

vi. To install the system as web based platform that would enable users to upload images,

pose questions and get the results in form of interpretation in real time.

1.4 Project Scope

The VisiHealth AI project scope can be explained by the following statement: the project aims at the development of a clear and transparent AI model in medical diagnostics. The system is a combination of the deep learning feature extraction and knowledge retrieval in order to give context aware responses.

Inclusions:

Multimodal Processing: Training a CNN and fine tuning BioBERT.

Dataset Use: The dataset we are going to use is the English data of the SLAKE dataset which consists of images, QA pairs, segmentation masks, and knowledge graph triplets.

Explainability Features: Region of interest (ROI) detection and template based rationale generation.

Web Application: This is a Web based interface which is developed and allows users to interact with the model without any local installation.

Exclusions:

Multi-Language Support: The system will specialize in the English QA pairs. While SLAKE is bilingual, this project restricts its scope to the English subset.

Clinical Deployment: The project is research and support prototype, the project will not entail extensive deployment in active clinical operations or clinical trials.

Large-Scale Pre-training: Fine-tuning of enormous pre-trained domain specific models will be done because of the constraints of resources, it will not be possible to train extremely large foundational models directly.

Target Users:

i. Radiologists: To support evidence based decision making by offering second opinion advice.

ii. Medical Students: To be used as a learning tool connecting the visual conclusions and theoretical medical information.

iii. Healthcare Researchers: Healthcare medical imaging has the potential to use Explainable

AI (XAI).

Project Deliverables:

An effective Med VQA prototype which will answer medical related queries.

A built in module giving text based clarification.

An online web interface to pictures upload and responses.

Chapter 2: Literature Review

2. Literature Review

2.1 Overview

Medical Visual Question Answering (Med-VQA) lies at the intersection of Computer Vision (CV) and Natural Language Processing (NLP), enabling AI systems to analyze medical images and respond to clinical questions. Unlike general-domain VQA, Med-VQA must handle anatomical precision, specialized medical terminology, and the high stakes of clinical decision-making. As a result, explainability, trust, and reasoning play a far greater role.

Recent advancements in Med-VQA focus not only on improving accuracy but also on integrating Explainable AI (XAI) techniques, such as visual grounding and knowledge-based reasoning. With the growing need for transparent medical AI systems, research has shifted toward models that justify their decisions through interpretable evidence. The proposed VisiHealth AI aligns with this direction by combining CNN-based feature extraction, BioBERT embeddings, Region of Interest (ROI) localization, and Knowledge Graph (KG)-driven rationale generation, as also emphasized in the project's methodology section

.

2.2 Deep Learning in Medical Image Analysis

Before multimodal systems emerged, medical AI research primarily focused on analyzing visual data through deep learning. These foundations serve as the "visual encoder" in Med-VQA systems.

2.2.1 Convolutional Neural Networks (CNNs)

CNNs remain the dominant architecture for extracting structured information from medical scans.

- **ResNet** (He et al. [6]) introduced residual connections, enabling deeper and more stable models. ResNet-50/101 are widely used in radiology due to their ability to detect subtle pathologies.

- **VGGNet** (Simonyan & Zisserman [7]) uses simple stacked convolutions but generates powerful representations, making it a common choice for transfer learning.

- **DenseNet**, used in QCR [17], improves gradient flow and feature reuse, which benefits training on small datasets.

These CNNs produce global image features but lack explicit localization of abnormalities—leading to further research into segmentation-based solutions.

2.2.2 Semantic Segmentation and ROI Detection

Precise localization of anatomical structures is crucial in clinical applications. U-Net (Ronneberger et al. [8]) became a landmark model for medical image segmentation due to its encoder–decoder design and skip connections.

In Med-VQA, ROI detection helps ensure that the system's answer is grounded in anatomically relevant regions. The SLAKE dataset used in VisiHealth AI provides segmentation masks for 39 organs and 12 diseases, which enables models to identify and highlight meaningful image areas before reasoning or answering questions. This segmentation-driven grounding is also emphasized directly in the proposal methodology

2.3 Evolution of Med-VQA Architectures

2.3.1 Generation 1: Joint Embedding Models

The earliest Med-VQA models used CNNs for images and LSTMs or Bag-of-Words models for questions. These features were fused through concatenation or bilinear pooling. Although simple, these methods struggled with:

- complex clinical terms,
- reasoning requirements, and
- spatial understanding.

They also lacked explainability.

2.3.2 Generation 2: Attention-Based Models

Attention mechanisms improved performance by highlighting important regions or words. Examples include:

- **Stacked Attention Networks (SAN)** [10]
- **Bilinear Attention Networks (BAN)**

However, these models often suffered from language bias predicting answers based on statistical patterns rather than true visual understanding. Heatmaps generated through attention did not provide clinically precise ROI evidence.

2.3.3 Generation 3: Transformer-Based Models

Transformers transformed both NLP and CV research.

- **BERT** [12] enabled contextual question embeddings.
- **BioBERT** [13], trained on biomedical texts, significantly improves understanding of clinical terminology.
- **Vision Transformers (ViT)** brought self-attention into imaging but require large amounts of training data.

Modern multimodal Transformers combine ViT with BERT/BioBERT to perform cross-modal reasoning. VisiHealth AI adopts this generation's principles through BioBERT text encoding and a fusion module that integrates image and KG features for more robust inference.

2.4 Knowledge-Enhanced Reasoning

2.4.1 Role of Knowledge Graphs

Traditional Med-VQA models cannot answer questions requiring medical expertise beyond what is visually detectable. Knowledge Graphs (KGs), represented as triplets (entity–relation–fact), support real-world medical reasoning.

Example:
(Cardiomegaly → associated with → Enlarged heart)

KDs enable:

- multi-hop reasoning,
- retrieval of disease–symptom relationships,
- more structured explanations.

2.4.2 SLAKE Dataset

SLAKE provides:

- radiology images,
- QA pairs,
- segmentation masks, and
- **2,600+ KG triplets**.

This combination uniquely supports explainability-oriented models. Because VisiHealth AI uses ROI segmentation and KG retrieval for rationale generation, SLAKE is particularly well suited to its architecture, as outlined in the proposal's dataset justification section

2.5 Explainable AI (XAI) in Med-VQA

Explainability is critical for medical deployment. Even accurate systems may be rejected by clinicians if they cannot justify their decisions.

2.5.1 Visual Explainability

Grad-CAM [14] highlights influential pixel regions for a model's prediction. While widely used, traditional heatmaps:

- lack precision,
- sometimes highlight irrelevant artifacts,
- cannot replace segmentation-based anatomical ROI masks.

VisiHealth AI therefore uses segmentation masks for clear, clinically grounded visual evidence.

2.5.2 Textual Rationales

Research (Vollmer et al. [16]) shows clinicians prefer textual explanations that clarify *why* a particular answer was chosen. Rationale generation offers transparency by linking:

- detected ROIs,
- KG facts, and
- predicted answers.

The VisiHealth AI proposal implements template-based rationale generation using retrieved KG triplets and localized image regions, ensuring human-readable explanations such as:

> "Detected enlargement in the left lung region. KG links this with Pneumonia. Therefore, answer = Yes."

Such reasoning chains improve user trust and educational value.

2.6 Comparative Analysis of Representative Models

Feature / Model	SAN (2016)	MEVF (2021)	QCR (2022)	VisiHealth AI
Visual Encoder	VGG	MAML	DenseNet	**Custom CNN + ROI masks**
Text Encoder	LSTM	BERT	BERT/BioBERT	**Fine-tuned BioBERT**
Fusion Strategy	Attention	Bilinear pooling	Cross-modal	**Knowledge-aware fusion**
Knowledge Integration	No	No	No	**Yes (KG retrieval)**
Visual Grounding	Weak	Weak	None	**Strong (Segmentation masks)**
Explainability	Low	Medium	Medium	**High (ROI + KG rationale)**

This comparison highlights that while accuracy has improved across generations, **explainability** remains limited in most frameworks. VisiHealth AI addresses this gap by integrating ROI detection and KG-based reasoning—features not present in earlier models.

2.7 Additional Considerations in Med-VQA Research

2.7.1 Dataset Challenges

Medical datasets are often small due to privacy concerns. SLAKE mitigates this with rich annotations, though the proposal's feasibility study acknowledges the risk of overfitting due

to its small size

Techniques such as data augmentation, dropout, and early stopping are therefore essential.

2.7.2 Fusion Techniques

Fusion is a core aspect of Med-VQA:

- **Early fusion**: simple concatenation
- **Bilinear pooling**: captures multiplicative interactions
- **Cross-modal attention**: deeper image–text alignment
- **Knowledge-aware fusion**: integrates external medical reasoning

VisiHealth AI employs a hybrid approach that combines CNN–BioBERT embeddings with KG retrieval.

2.7.3 Evaluation Metrics

Med-VQA systems are evaluated through:

- **Accuracy** (categorical answers),
- **BLEU/METEOR** (textual answers),
- **IoU** (grounding via segmentation masks),
- **Human evaluation** (explanation quality).

As XAI becomes more important, rationale evaluation is becoming increasingly standardized.

2.8 Summary and Research Gaps

The literature reveals several persistent challenges in Med-VQA:

1. Limited Visual Grounding

Attention-based heatmaps are not precise enough for clinical reliability.

2. Lack of Actionable Explanations

Most models provide answers without reasoning, making them unsuitable for medical decision support.

3. Poor Integration of External Knowledge

KG-based reasoning is rare due to dataset limitations.

4. Small Dataset Constraints

Models risk overfitting, as acknowledged in the project's feasibility analysis, and require strong regularization strategies.

VisiHealth AI: Addressing Existing Gaps

VisiHealth AI offers a solution by:

- adopting ROI segmentation for accurate visual grounding,
- integrating KG triplets for medically-sound reasoning,
- generating template-based textual rationales, and
- combining multimodal deep learning with structured knowledge retrieval.

This positions the system as a modern, explainability-first Med-VQA architecture suited for both medical education and decision support.