K-PathVQA: Knowledge-Aware Multimodal Representation for Pathology Visual Question Answering

Usman Naseem, *Student Member, IEEE*, Matloob Khushi, Adam G. Dunn and Jinman Kim, *Member, IEEE*

***Abstract*— Pathology imaging is routinely used to detect the underlying effects and causes of diseases or injuries. Pathology visual question answering (PathVQA) aims to enable computers to answer questions about clinical visual findings from pathology images. Prior work on PathVQA has focused on directly analyzing the image content using conventional pretrained encoders without utilizing relevant external information when the image content is inadequate. In this paper, we present a knowledge-driven PathVQA (K-PathVQA), which uses a medical knowledge graph (KG) from a complementary external structured knowledge base to infer answers for the PathVQA task. K-PathVQA improves the question representation with external medical knowledge and then aggregates vision, language, and knowledge embeddings to learn a joint knowledge-image-question representation. Our experiments using a publicly available PathVQA dataset showed that our K-PathVQA outperformed the best baseline method with an increase of 4.15% in accuracy for the overall task, an increase of 4.40% in open-ended question type and an absolute increase of 1.03% in closed-ended question types. Ablation testing shows the impact of each of the contributions. Generalizability of the method is demonstrated with a separate medical VQA dataset.**

***Index Terms*— Pathology Images, Medical Visual Question Answering, Multimodal Representation**

I. Introduction

MEDICAL visual question answering (MedVQA) task aims to correctly answer a question related to a medical image and related text-based clinical reports and can help in diagnosis and training [1], [2]. Pathology visual question answering (PathVQA) is a domain-specific MedVQA task, specifically for pathology images, and involves comprehending medical-related concepts while considering the pathology images and text information.

Existing works on PathVQA focus on individually computing a question representation using language models and image features using pre-trained convolutional neural networks (CNN)-based models; these features are then combined and forwarded to another separate network to train a model [3]–[8]. However, all these works relied on language models that were trained for the general field, and therefore did not account for the large differences in the language model that is necessary for the medical field, e.g., with the use of acronyms and domain-specific medical terminologies. Therefore, existing PathVQA benchmarks often performed well only when the visual contents were the distinct feature of the query. Despite this, with PathVQA, it requires the use of medical knowledge other than the images/text information to answer difficult compositional questions involving inquiries such as "*the location of a disease*", "*the treatment of a disease*", "*the cause and symptom of a disease*", or "*the functionality of an organ*", as exemplified in Fig. 1.

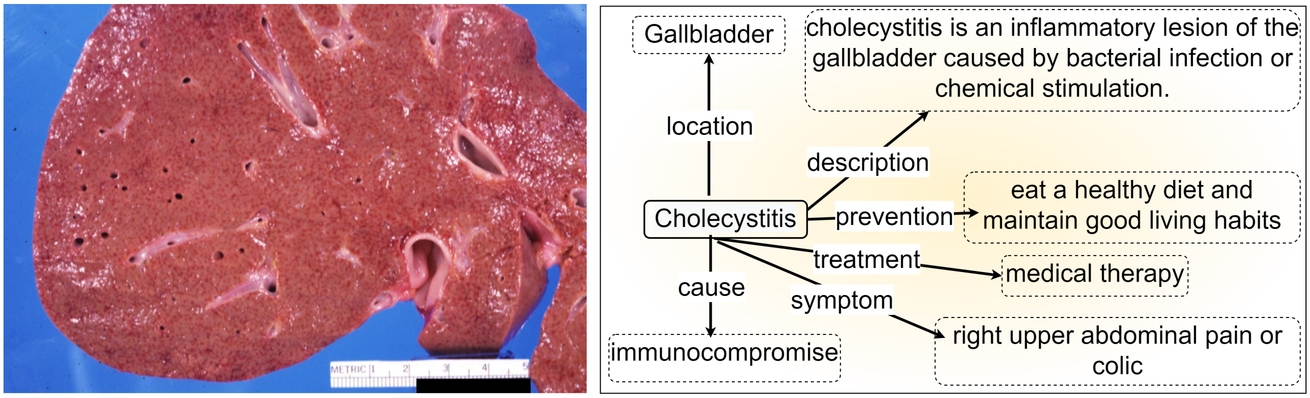
To address this, PathVQA needs to incorporate domain-specific knowledge bases (KBs) [9] with graphical representation where entities are represented by nodes and directed edges connecting the nodes represent their relationships.e.g., *(Gastrointestinal, Location, Stomach)*. Apart from reliance on non-domain specific language models, these existing PathVQA approaches [3]–[8], [10]–[12] also relied exclusively on the information within the image, and this caused wide failures when the visual content of a medical image was insufficient to answer the questions. There is a need for external knowledge other than the image content to answer complex questions. Recently, Naseem et al. [8] presented Trap-VQA that fused the image and text features extracted using ResNet [13] and BioELMO [14], a domain-specific language model respectively to the transformers' encoder layers for PathVQA. Trap-VQA increased the performance compared to the previous models; however, it was unable to answer questions when the image content was limited and required external medical knowledge.

This study presents a novel PathVQA model that effectively integrates image, question, and knowledge representations, that captures knowledge-image-question-specific interactions. Our model goes beyond conventional approaches for PathVQA by incorporating external medical knowledge into the PathVQA task. This is achieved by fusing visual and question features with representations of external medical knowledge.

To realize this integration, we propose a comprehensive framework consisting of image, question, and knowledge representation modules. The Image-Question representation module enhances both the image and question representations

This project was supported in part by ARC DP200103748.

U. Naseem, and J. Kim are with the School of Computer Science, The University of Sydney, Australia. M. Khushi is with the Department of Computer Science, Brunel University London, London, UK. AG. Dunn is with the School of Medical Sciences, The University of Sydney, Australia. (e-mails: {usman.naseem, adam.dunn, jinman.kim}@sydney.edu.au; matloob.khushi@brunel.ac.uk).


METRIC 1 2 3 4
Gallbladder
location
cholecystitis is an inflammatory lesion of the gallbladder caused by bacterial infection or chemical stimulation.
description
eat a healthy diet and maintain good living habits
Cholecystitis
prevention
treatment
medical therapy
cause
symptom
immunocompromise
right upper abdominal pain or colic

Q: What is thickened and fibrotic due to chronic *cholecystitis*?
A: the wall of the *gallbladder*

A snippet of a medical knowledge graph for "*Cholecystitis*" disease

Fig. 1. Examples of an image-question pair from the PathVQA dataset that requires external medical knowledge (right) to correctly answer a question (related to "Cholecystitis") associated with a pathology image (left).

by mutually enriching contextual information. Additionally, the Knowledge-Question representation module leverages a medical knowledge graph embedding to incorporate relevant external information into the question embedding.

The final step involves aggregating the representations from the Image-Question and Knowledge-Question modules using the Knowledge-Image-Question representation module. This module combines the enhanced image and question representations with the incorporated external medical knowledge, resulting in a comprehensive representation that captures the intricate interactions between knowledge, image, and question. Our key contributions are as follows:

- We present a novel end-to-end trainable method, namely knowledge-PathVQA (K-PathVQA), that incorporates medical knowledge to PathVQA by introducing a knowledge graph constructed from complementary medical factual knowledge from external structured contents to answer questions that require information beyond visual content.
- We introduce a multimodal representation that jointly learns a Knowledge-Image-Question representation that aggregates the representations from Image-Question and Knowledge-Question modules without requiring additional knowledge annotations or search queries for PathVQA.
- Our experiments demonstrated that incorporating medical knowledge into our model outperformed the previous state-of-the-art methods on the benchmark PathVQA dataset and is also generalizable on another medical VQA dataset.

II. Related Work

A. Medical Visual Question Answering

Prior studies on MedVQA [22]–[27] have been an adaptation of methods developed for general-domain VQA models such as Bilinear Attention Networks (BAN) [15], Stacked Attention Networks (SAN) [16], and Multi-modal Compact Bilinear (MCB) [17], where attention mechanism and bilinear pooling schemes are applied to capture cross-modal feature fusion which captures the textual and visual relationship. SAN locates question-related visual regions using multi-step inference by extending the attention mechanism. The derived features are used in the classifier to predict the answers. MCB, on the other hand, introduced a multi-modal compact bilinear pooling method that lowers the computation of feature fusion by projecting the outer product to a lower dimensional space. BAN is derived from the fusion of bilinear multi-modal in MCB and uses the low-rank bilinear pooling to lower the rank of weight, the outer product of several model vectors, to lower the computation cost. Other VQA methods include Multimodal Factorized High-order (MFH), and Multi-modal factorized bilinear (MFB) [19], which are proposed to reduce computational cost and are built on a similar concept of generating the bilinear pooling of two vectors computationally efficient by decomposing the outer product projection matrix. These methods also dominated the MedVQA's ImageCLEF challenges. Pre-trained CNN-based methods such as ResNet [13] or VGGNet [28] are usually used to obtain visual features. For textual features, recurrent neural networks (RNNs) [29] and transformer-based language models such as Bidirectional Encoder Representations from Transformers (BERT) [30] are used to derive text-based features.

Using ResNet and RNN for deriving features from image and text, Peng et al. [26] adapted MFH to fuse the image-text features in the first ImageCLEF challenge. Inception-Resnet and BiLSTM were applied by Zhou et al. [27] to represent visual and text features that are then combined to classify answers. Abacha et al. [23] applied VGG and LSTM to obtain visual, and text features before fusing the text and image features using SAN. The best method [25] in the ImageCLEF 2019 challenge (2nd edition) used BERT and VGG for text and visual features that are then fed to MFB for fusion for prediction. The best method (AIML) [31] in the third ImageCLEF 2020 competition first classified the question by separating the question types as open-ended or closed-ended and consequently modified the VQA to a multi-task image classification task.

However, differences between medical and general questions mean that applying general domain VQA methods to medical questions does not yield optimal results. Examples of differences between the problems include limited datasets of medical image and text data for training, specific need to be able to detect and classify anatomical and functional structures in medical images, and uniqueness of the medical knowledge

TABLE I
COMPARISON OF ADVANTAGES AND DISADVANTAGES OF PREVIOUS METHODS

Method	Advantages	Disadvantages
BAN [15]	Can capture cross-modal feature fusion. Can learn the relationship between text and image features	Computationally expensive. Unable to capture long-range dependencies between the image and the question. Unable to generalize to new data.
SAN [16]	Can locate question-related visual regions. Can use multi-step inference to improve accuracy.	Computationally expensive. Sensitive to the quality of the images. Unable to generalize to new data.
MCB [17]	Reduces computation of feature fusion. Computationally less expensive than BAN and SAN	Unable to capture long-range dependencies between the image and the question. Unable to handle multipel objects in an image. Unable to handle opn-ended questions.
MFH [18] and MFB [19]	Reduce computational cost. Efficiently generate bilinear pooling of two vectors	Prone to overfitting. Limited generalizibility Sensitive to noisy data and require large amount of training data.
MEVF [20]	Robust to nouse in the input images. Efficient training	Coputationally expensive. Can not generalize to other tasks.
CGMVQA [21]	Integrates a classifier and a generator. Uses a transformer's multi-head self-attention	Unable to handle open-ended questions. Require large amount of trainign data. Dependency on pre-defined anaswer classes.
CMSSL [4] and 3LA [5]	Jointly captures image and language features for PathVQA. Self-supervised pretraining and VQA fine-tuning. Excludes noisy self-supervised samples	Require large amount of paried image-text data. Computational complex. May lead to overfitting.
MedFuseNet [6]	Learnt essential components of a medical image and effectively answered medVQA, including PathVQA	Rely on external data for transfer learning. Metadata is not fully used.
MMQ [7]	Enhances metadata by auto-annotation Handles noisy labels in the training stage	Unable to capture the high and low-level interactions of data
Trap-VQA [8]	Uses a transformer-based method for PathVQA	Unable to answer medical questions when the image contents are insufficient

such as specialised medical terminologies.

Researchers in the MedVQA community have presented several methods that are designed for medical images. For instance, Nguyen et al. [20] introduced a Mixture of Enhanced Visual Features (MEVF) to overcome limited medical data by initializing the visual feature model weights. Extending on the previous study (i.e., MEVF), Zhan et al. [32] introduced two new modules that are conditioned on a question and type reasoning that use the MEVF visual backbone to train VQA models' reasoning skills. In another study, Ren and Zhou [21] introduced a classification and generating approach for MedVQA (CGMVQA) that integrates a classifier and a generator and uses a transformer's multi-head self-attention; it outperformed the method used by a VQA-Med-2019 challenge winner.

Furthermore, despite the effectiveness of these methods on various medical images for VQA tasks, their applicability for PathVQA has not been tested. Also, current MedVQA methods have shown promising results in answering questions related to medical images. However, these methods are limited in incorporating medical knowledge to answer questions accurately, and therefore often fails to understand the context of the question and the medical images, and thus leading to incorrect answers. For instance, if a MedVQA model is asked to identify a skin lesion, it may provide an answer based solely on the appearance of the lesion. In contrast, a doctor may consider the patient's medical history, symptoms, and other relevant information to make a more accurate diagnosis.

B. Pathology Visual Question Answering

Several works have been developed to specifically address PathVQA tasks. For example, He et al. [3] proposed leveraging cross-modal self-supervised learning to jointly capture image and language features for PathVQA tasks. In another study, He et al. [5] proposed a three-level approach for optimization that performed self-supervised pretraining and VQA fine-tuning to capture image and language features jointly and automatically excluded noisy self-supervised samples from pretraining. Sharma et al. [6] introduced MedFuseNet, an attention-based multimodal-based method that learnt essential components of a medical image and effectively answered medVQA, including PathVQA. Another study by Do et al. [7] introduced multiple meta-model quantifying (MMQ), which enhances meta-data by auto-annotation and handles noisy labels in the training stage by using the uncertainty of predicted results during the meta-agnostic process to create meta-models with strong features for PathVQA. Recently, Naseem et al. [8] proposed a transformer-based method (Trap-VQA) for PathVQA. In Trap-VQA, image and textual features are fused to transformers' encoder layers for the final prediction.

One of the main limitations of the current PathVQA methods are their inability to accurately answer medical questions when the image contents are insufficient to provide an answer. This limitation can be attributed to the fact that current PathVQA models mainly rely on visual features extracted from the images without considering the vast amount of medical knowledge required to answer medical questions accurately (See Table I for detailed list of advantages and disadvantages of previous methods).

Medical knowledge is crucial in PathVQA, as it can provide additional information that may not be visible in the image alone. For example, a medical expert may be able to infer certain details about a patient's condition based on their medical history or other related factors that are not apparent in the image. By incorporating medical knowledge into the PathVQA task, a model can make more accurate predictions and provide improved insights for doctors.

Therefore, by leveraging external medical knowledge facts, our K-PathVQA model aims to address this limitation of current PathVQA methods and provide more accurate answers to medical questions. By doing so, we hope to improve the utility of PathVQA to aid doctor's make informed decisions.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/jbhi.2023.3294249,

C. Language Models

Language models have been widely used in various natural language processing (NLP) tasks, such as speech recognition, machine translation, and sentiment analysis. Two types of language models have been widely studied: general language models and domain-specific language models.

General language models [33], [34] are pre-trained on large amounts of text data from a variety of sources and can generate coherent and contextually relevant text. They have been shown to achieve state-of-the-art performance in various NLP tasks, including text classification and language generation [35]. However, general language models may not perform as well in domain-specific tasks as they lack the domain-specific knowledge necessary to generate accurate predictions [36], [37].

To address this issue, domain-specific language models have been developed and used in various applications. For example, BERT [33] is a domain-specific language model that has been pre-trained on text data from the biomedical domain and has been shown to improve performance on biomedical NLP tasks [38], [39]. Similarly, PHS-BERT [37] is a domain-specific language model trained on social media textual data and shown to outperform general language models on tasks related to public health surveillance on social media.

Despite the success of general and domain-specific language models in various NLP tasks, their application to PathVQA remains limited.

III. Method

Problem Formulation: Given a pathology image V $\in$ V associated with a related clinical question Q $\in$ Q and a knowledge graph G, we aim to produce an answer $\hat{A} \in A$. The predicted answer $\hat{A}$ of the proposed model is mathematically expressed as:

$$\hat{A} = \arg\max_{A\in\mathcal{A}} p_\theta(A|V,Q,\mathcal{G}) \tag{1}$$

where θ symbolises the parameters of the model p that needs to be trained. To correctly predict the correct, our goal is to learn a joint representation $z \in R^{d_z}$ of V, Q, and G such that:

$$A^* = \hat{A} = \arg\max_{a\in\mathcal{A}} p_\theta(a|z) \tag{2}$$

where A^* is the true answer. The dimension of the joint space z is denoted by the hyperparameter d_z, which is chosen as a result of a trade-off between representation capability and computational cost.

Overview of the proposed method: Our proposed Knowledge-Aware Multimodal Representation for Pathology Visual Question Answering (K-PathVQA) comprises of three main modules as illustrated in Fig. 2: (i) Input layer, which comprises an image representation, a question representation, and a knowledge representation; (ii) Image-Question representation where we generate image-attended question features; (iii) Knowledge-Question representation which generated question features conditioned on knowledge embeddings, (iv) Knowledge-Image-Question representation module where we aggregate outputs to predict an answer. Each of these modules is discussed below in detail.

A. Input Layer

The input layer of K-PathVQA consists of an image representation, a question representation, and a knowledge representation. Below we explain each of these in detail.

Image representation: To extract image features ($v \in \mathbb{R}^{d_v}$), we adopted a transfer-learning approach by using a pre-trained ResNet50. We reshaped an image to match the shape of ResNet50 (224, 224, 3) - an input image with a height of 224 pixels, a width of 224 pixels, and three color channels (RGB) and output features of 2048. We removed the last three fully connected layers and kept the output from the last averaged pooled layer to obtain image features.

Question representation: Given a question Q consisting of n_T tokens, we use pre-trained biomedical version of ELMo [40] language model i.e., BioELMO [14] to generate question representation $q \in \mathbb{R}^{n_T x d_q}$.

Knowledge representation: To obtain knowledge representation, we leveraged the medical knowledge graph [41] centred on organs and related disorders. A set of 52.6K triplets *(head, relation, tail)* comprising medical facts was collected from OWNThink[1], a large knowledge database based on Wikipedia in the medical knowledge graph we used. In the KG, edges represent the relationship between the entities, such as between a function and its treatment, while the nodes are entities, such as organs or diseases. Triplets are then made of two entities and the relation between them. We refined triplets so that an organ's function or body system must be described in the triplets that refer to it, and the symptoms, locations, causes, and methods of treatment or prevention of disease must all be described in triplets for that disease. We used 2,603 triplets in the English language. Table II shows examples of the triples and their relationship between entities in medical knowledge (i.e., SLAKE) used in our experiments.

TABLE II
Examples of a Medical Knowledge Graph.

Organ	<Heart, Function, Promote blood flow> <Kidney, Belong to, Urinary System> <Duodenum, Length, 20-25cm>
Disease	<Pnuemonia, Location, Lung> <Lung Cancer, Cause, Smoke> <Brain Tumor, Symptom, visual impairment>

Our domain-specific knowledge representation consists of knowledge retrieval and knowledge embedding generation. In knowledge retrieval, we first extracted the entities, i.e., nouns, from the questions using SpaCy and then mapped those extracted entities with entities present in our medical KG. After this step, during knowledge embedding generation, we feed these entities with their relation to biomedical version of BERT, i.e., BioBERT [38], a domain-specific language model to generate knowledge representation $k \in \mathbb{R}^{n_T x d_q}$.

B. Image-Question representation

To learn question q and image v representations jointly, we extract image-question features VQ using our image-question encoder module. Our image-question encoder module

[1] https://www.ownthink.com/knowledge.html

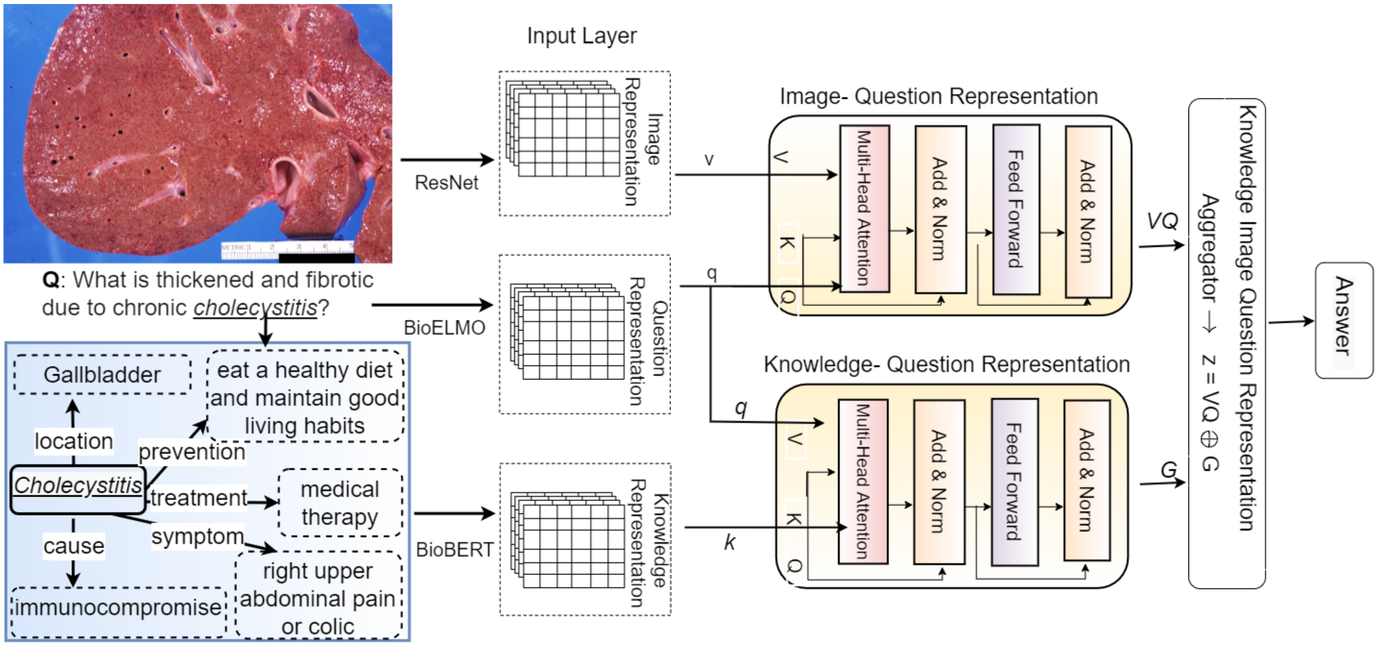


Fig. 2. Overall architecture of the proposed K-PathVQA

is mainly based on two transformer encoder layers to image-question features. Given input question features q and image features v, Our model generates a VQ-representation as a final result.

The input to the first layer of the transformer encoder of the image-question encoder contains 'values' of image representations and 'queries' and 'keys' from question representations. We employed multi-head attention in conjunction with a scaled dot-product. As a result, we packed a set of v into a matrix V_v, and q into a matrix Q_q and K_q.

$$Att_1(Q_q, K_q, V_v) = softmax(\frac{Q_q.K_q^T}{\sqrt{d_k}}).V_v \tag{3}$$

The final output, VQ, of the second transformers' encoder block of the image-question encoder module, represents image-attended question features. The final image-attended question features represent high-level information of both image and question features. The hyper-parameters have been explained in Section IV-B (Experimental settings).

C. Knowledge-Question representation

We present the knowledge-question representation module to extract question features conditioned on knowledge representations. Knowledge-question representation module performs knowledge-conditioned question attention by feeding the question features q and the KG representations k to the two-layered encoder layers of the original transformer. As a result, the model can incorporate domain-specific, i.e., medical knowledge, to the question and improve the understanding of the question related to the knowledge contained in the KG. The knowledge-question representation module outputs the knowledge-enriched question representation referred as G.

The input to the first layer of the transformer's encoder of knowledge-question's encoder contains 'values' of question representations and 'queries' and 'keys' from KG representations. We employed multi-head attention in conjunction with a scaled dot-product. As a result, we packed a set of q into a matrix V_q, and **k** into a matrix Q_G and K_G.

$$Att_2(Q_q, K_G, V_G) = softmax(\frac{Q_q.K_G^T}{\sqrt{d_k}}).V_G \tag{4}$$

The final output, G, is extracted from the second transformer's encoder block of the knowledge-question encoder module, representing knowledge-attended question features. The final set of knowledge-attended question features represents a new representation of the question, enhanced with domain-specific medical knowledge extracted from the KG.

D. Knowledge-Image-Question Representation module

We concatenate the outputs of the image-question and knowledge-question modules to produce a unified knowledge-image-question representation. We generate a vector that jointly represents the two features (VQ and G) and contains both higher and low-level interactions.

$$z = VQ \oplus G \tag{5}$$

where element-wise symbol $\oplus$ represents a vector's aggregation and (VQ and G) are the embeddings generated by image-question and knowledge-question representation modules. The output of the aggregator z is a joint knowledge-image-question representation which is then forwarded to predict the answer.

IV. Experiments

A. Dataset

We used a publicly available PathVQA dataset [3] containing 4,998 images and 32,799 questions (Table III). The questions were categorized into 7 groups: 6 groups are open-ended questions (what (40.9%), where (4.0%), when (0.9%), whose (0.6%), how (3.0%), how much/how many (0.9%)) and one group is closed-ended (yes/no (49.8%)) questions. We used the standard training, validation, and test sets comparable with earlier PathVQA works [3], [8] to evaluate our model.

TABLE III
DATA STATISTICS: DIVISION OF TRAINING, VALIDATION AND TEST DATA

Dataset	Training data	Validation data	Test data
# images	2,499	1,499	1,000
# QA pairs	17,325	9,462	6,012

B. Experimental settings

1) Implementation settings:

Question representation: We investigated different language models to generate question representation (e.g., BioELMO, BERT, and BioBERT). From our empirical experiments, we found that embedding generated by BioELMO performed best as compared to others in our model. Therefore, the input questions were embedded using a pre-trained BioELMO language model. Each word was represented by a **768**-D word embedding (d_q = **768**). We set (n_T = 16) so that every question was split into 16-tokens beginning with a `[CLS]` token and ending with a `[SEP]` token. These 16-tokens were then represented as one-hot encoding. Appendix contains experimental results (Table X).

Image representation: We tested different pre-trained models to extract image features (e.g., VGG19, InceptNet, DenseNet, and ResNet. We empirically found that ResNet50 performed the best compared to others. The input images are embedded using pre-trained ResNet50. Appendix contains experimental results (Table X).

Knowledge graph representation: In our experiments, we used different methods (e.g., BERT, BioBERT, and BioELMo) to generate knowledge embeddings. We found that embedding generated by BioBERT worked best in our model (Table VII).

Transformer Encoder layers: In both of our encoder modules, we tested with a different number of transformers' encoder layers. We identified that using two transformers' encoder layers performs better than others (Table VIII). Our transformer encoder modules include two layers of transformer blocks, six layers with eight attention heads, and concatenated vector of both encoders results in a vector with dimension 512.

2) Hyperparameters settings: We used the Adam optimizer with a learning rate scheduler according to the formula in [42] with $\beta 1$ = 0.9, $\beta 2$ = 0.98 and $\epsilon = 10^9$ to train our model. A batch size of 64 for 50 epochs, 4000 warmup steps, and the grid-search optimization was used to find the best settings. We evaluated different layers of transformers and adopted various image and question feature extraction methods. All results are reported using the accuracy evaluation metric, the standard metric used in similar previous works [3], [8].

C. Baselines

We evaluated the performance of our model with PathVQA methods as well as general VQA and MedVQA methods, including the state-of-the-art PathVQA methods. Details of our comparison methods are as follows.

- **General VQA methods:**
  - Bilinear Attention Networks (BAN) [15] encodes visual and language features using a Gated Recurrent Unit and a Faster R-CNN. It captures bilinear attention distributions employing BAN and approximates the bilinear interaction between question and image representations by employing the low-rank approximations.
  - Multi-modal Compact Bilinear (MCB) [17]: A CNN is employed to encode the image, while an LSTM is used for textual features. An MCB pooling system is applied to predict the answer via an attention method.
  - Stacked Attention Networks (SAN) [16]: With CNN and LSTM to encode textual and visual features, the SAN incorporates a multiple-layer attention mechanism that iteratively infers the answer by repeatedly querying an image to identify the relevant image region.
  - Multi-modal factorized bilinear (MFB) [19] encodes visual and textual features using CNN and LSTM and employs MFB pooling to integrate textual and visual features.
- **Vision language methods:** We also compared our results with current state-of-the-art vision language pre-trained models such as LXMERT [43], VisualBERT [44] and UniTER [45] to fuse image and the textual features obtained using CNN and LSTM.
- **Medical/Pathology VQA methods:**
  - Mixture of Enhanced Visual Features (MEVF) [20] derives visual and textual features using CNN and LSTM and employs a MEVF with SAN and BAN to integrate visual and textual features for MedVQA.
  - Cross-modal self-supervised learning (CMSSL) [4] identifies and disregards noisy self-supervised samples to train PathVQA with visual and textual features.
  - Multiple meta-model quantifying (MMQ) [7] increases the meta-data by auto-annotation, deals with noisy labels in the training stage by utilising the uncertainty of predicted results during the meta-agnostic method and generates meta-models with robust features for MedVQA including PathVQA.
  - MedFuseNet [6] is an attention-based method that learns the essential features of a medical image and efficiently answers the questions. MedFuseNet has been evaluated on medical images, including PathVQA.
  - Trap-VQA [8], a recent state-of-the-art method for PathVQA that fuses the image and text features to the transformers' encoder layers for PathVQA.

D. Results

Table IV compares the performance of K-PathVQA to the state-of-the-art baseline methods. Compared to the best performing baseline method (Trap-VQA), K-PathVQA resulted in an increase of 4.15% in performance for the overall task (68.97%), an increase of 4.40% in open-ended question type (42.12%) and an increase of 1.03% in closed-ended question types (94.60). The results demonstrate that medical/pathology VQA methods like MEVF+BAN and MEVF+SAN outperformed the general VQA methods, including BAN and SAN methods; however, the performance is less compared to transformer-based methods like LXMERT, VisualBERT, UniTER, LXMERT+CMSSL and Trap-VQA. These results highlight the importance of transformers to capture global relationships. Our experiments in Table IV show that our K-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/jbhi.2023.3294249, IEEE Journal of Biomedical and Health Informatics

TABLE IV

COMPARISON: K-PATHVQA (PROPOSED) V/S THE BASELINES.

Task	Model	Overall	Open-ended	Close-ended
General VQA	MFB [19]	39.85	20.15	53.77
	SAN [16]	42.43	23.40	59.40
	MCB [17]	57.04	29.03	57.60
	BAN [15]	55.10	33.50	68.20
MedVQA	MEVF +SAN [20]	57.10	25.87	86.90
	MEVF +BAN [20]	57.90	26.75	87.50
Vision Language	LXMERT [43]	60.00	35.33	83.00
	VisualBERT [44]	60.08	33.03	86.99
	UniTER [45]	60.33	33.79	87.70
PathVQA	MMQ [7]	48.80	13.40	84.00
	MedFuseNet [6]	38.10	15.80	63.60
	LXMERT+CMSSL [4]	60.10	34.50	87.10
	BAN+CMSSL [4]	58.40	33.50	87.20
	Trap-VQA [8]	64.82	37.72	93.57
	K-PathVQA (proposed)	**68.97**	**42.12**	**94.60**

PathVQA outperforms the baselines and the state-of-the-art models for both open-ended and close-ended categories. In the following sections, we discuss individual modules that contribute to the overall performance.

E. Ablation analysis

To investigate the contributions of individual modules in our K-PathVQA, we conducted three ablation analyses.

- **Module wise Comparison:** We compared the ablated instances of K-PathVQA relative to its complete form. Table V reports the overall accuracy in the following setting:
  - L: Only question features q are fed to the classifier.
  - VL: Only the outputs of the Image-Question representation module [V; Q] is concatenated and fed to the classifier.
  - KL: Only the output of the Knowledge-Question representation module G is fed to the classifier
  - K-PathVQA: the outputs of both Image-Question and Knowledge-Question modules are fused and fed to the decoder.

Comparison between L and KL instances demonstrate the importance of incorporating external knowledge. Adding the KL embeddings to the model led to a gain of 1.39% (overall), 1.07% (open-ended), and 4.48% (close-ended). As expected, the VL model outperforms the KL model, where most of the questions in the dataset are related to questions referring to what is visible in the image.

Effectiveness of using medical knowledge and knowledge-Image-Question representation: We performed experiments to demonstrate the effectiveness of the proposed knowledge-image-question representation module (Table VI). First, we incorporated the same medical knowledge representation used in our method to the best performing baseline (Trap-VQA). Then we removed the knowledge from our method and compared it with Trap-VQA. From the results in Table VI, we can see that incorporating knowledge to the Trap-VQA improves the performance by 1.38% for an overall task (66.20%), an increase of 1.88% in open-ended question type (39.60%) and an increase of 0.43% in closed-ended question types (94%) when compared to the original Trap-VQA i.e., without knowledge. We also observe from Table VI that even with the removal of the knowledge from our method, our model still achieved the best performance on all three tasks compared to the Trap-VQA. From these results, we can imply that both the knowledge and the proposed multimodal representation that jointly learns a knowledge-image-question representation adds to the overall performance for PathVQA.

TABLE V

ABLATION ANALYSIS

Model	Overall	Open-ended	Close-ended
L	55.90	21.46	88.00
KL	57.29	22.53	92.48
VL	65.07	40.02	93.12
KVL (Proposed)	**68.97**	**42.12**	**94.60**

TABLE VI

EFFECTIVENESS OF USING MEDICAL KNOWLEDGE AND KNOWLEDGE-IMAGE-QUESTION REPRESENTATION

Methods	Overall	Open-ended	Close-ended
Effectiveness of using a medical Knowledge			
Proposed	68.97	42.12	94.60
K-Trap-VQA	66.20	39.60	94.00
Effectiveness of Knowledge-Image-Question representation			
Proposed w/o K	65.07	40.02	93.12
Trap-VQA	64.82	37.72	93.57

Different knowledge type-wise Comparison: To analyze the effect of using different language models to generate knowledge embeddings, we replaced knowledge embeddings generated by different language models while keeping the same experimental settings as described in Section IV-B. Table VII shows that BioBERT resulted in the best performance in generating knowledge embeddings in all 3 tasks. We attribute BioBERT's performance to its domain-specific language model trained on biomedical data. We also observed that the performance of BERT on closed-ended question types is less compared to both (BioELMO and BioBERT) domain-specific language models. We postulate this due to the fact that BERT is trained on general domain corpus whereas both domain-specific language models are trained on biomedical corpus that helps to extract better text representation compared to BERT.

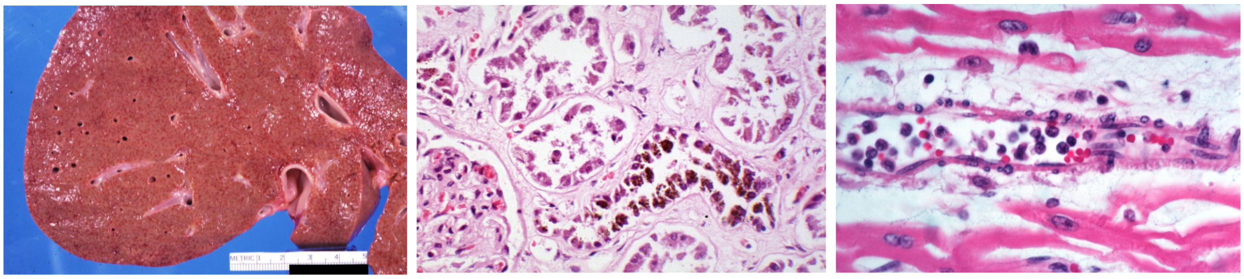
TABLE VII

COMPARISON WHEN TESTED ON USING DIFFERENT LANGUAGE MODELS TO GENERATE KNOWLEDGE REPRESENTATIONS

KG	Overall	Open-ended	Close-ended
BERT	68.94	40.82	86.42
BioBERT	68.97	42.12	94.60
BioELMO	68.74	38.57	92.92

Layer-wise Comparison: To analyze the effect of transformer layers in both Image-Question representation and Knowledge-Question representation modules, we tested our K-PathVQA with different transformer layers (ranging from 1 to 6). All experimental settings were the same as described in Section IV-B. The results are presented in Table VIII where in all 3 tasks, 2 layers of the transformer produced the best results. This drop in performance when using more than 2 transformer layers is attributed to the loss of features' information when more layers are added. These results demonstrate that using 2 layers of a transformer is the optimal number for getting better accuracy for PathVQA.

Generalizability of K-PathVQA: There is no other public pathology VQA dataset available. Therefore, to investigate the



Q: What is thickened and fibrotic due to chronic cholecystitis?
A: the wall of the gallbladder

VL: kidney,
Proposed: the wall of the gallbladder

(a)

Q: What caused by numerous blood transfusions?
A:hemosiderosis

VL: the red muscle ,
Proposed: hemosiderosis

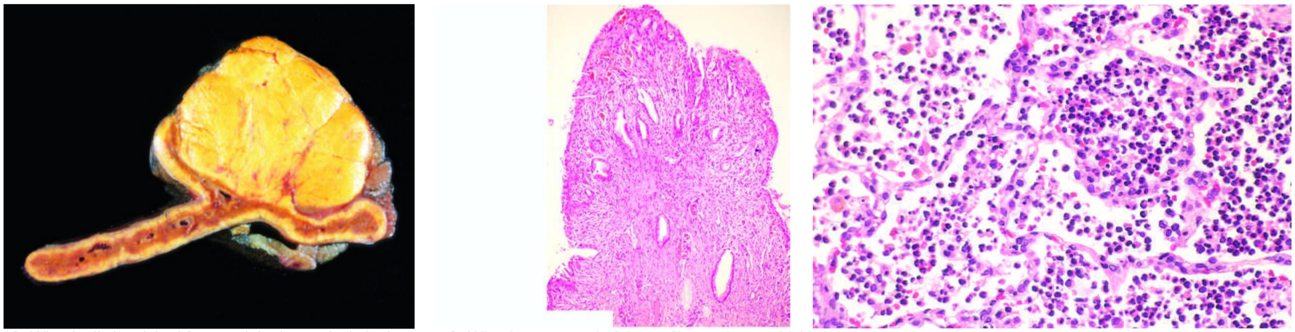
(b)

Q: What is rocky mountain spotted?
A:fever

VL: oral
Proposed: fever

(c)

Fig. 3. Qualitative results: Proposed model outperforms answers predicted by VL model. Answers that are correct are highlighted in green, while those that are wrong are highlighted in red.



Q: What is distinguished from nodular hyperplasia by its solitary, circumscribed nature?
A: adenoma
VL: necrosis,
Proposed: fibroma

(a)

Q: What is composed of dense fibrous tissue which shows nonspecific inflammation?
A: stromal core
VL: the deposits,
Proposed: storma

(b)

Q: What have not yet formed?
A: fibrin nets

VL: the red cells,
Proposed: the fibrous tissue

(c)

Fig. 4. Qualitative results: proposed model predicts answers of the same type as the ground-truth answer compared with the VL model on the question. Answers predicted by proposed model are indicated in blue, while answers predicted by VL model are shown in red.

TABLE VIII
COMPARISON WHEN TESTED ON DIFFERENT NUMBER OF LAYERS

No. of Layers	Overall	Open-ended	Close-ended
Layer = 1	67.87	39.41	91.66
Layer = 2	**68.97**	**42.12**	**94.60**
Layer = 3	66.72	41.98	94.15
Layer = 4	66.67	41.64	93.87
Layer = 5	65.90	41.75	86.28
Layer = 6	67.25	40.16	91.18

generalizability and the effectiveness of medical knowledge to other MedVQA datasets, we used SLAKE [41] – a MedVQA radiology images dataset, following a previous similar study [8]. SLAKE contains 642 images that contain a variety of modalities such as CT, MRI, and X-Ray, the coverage of body parts such as the head, neck, and chest, and 14,028 question pairs. We used the same experimental settings as used in [8] to train our model. In this experiment, we compared the results of K-PathVQA with the Trap-VQA which is the second-best method according to Table IV including the SOTA method used in [41]. Our results shows that K-PathVQA outperformed both SOTA results reported in [41] and Trap-VQA on all three question types (Table IX). We attribute this increase to the factual structured medical knowledge that helps our model to learn better question representation for questions related to radiology images. For example, to answer a question "*What organ belongs to the immune system*?", our method leverages the following factual knowledge <*Spleen, function, improve the body's immunity*> to retrieve the correct answer. These results indicate that leveraging a medical knowledge not only improves the performance on PathVQA dataset but also on other MedVQA dataset (i.e., SLAKE). This increase in performance on SLAKE dataset that contains radiology images shows that our model is generalizable and is not limited to only pathology images.

TABLE IX
ACCURACY (%) COMPARISON OF PROPOSED METHOD V/S THE BEST BASELINE (TRAP-VQA) ON OTHER MEDVQA (SLAKE) DATASETS. ↑ REPRESENTS THE INCREASE IN ACCURACY.

Type	Trap-VQA	K-PathVQA	↑ Accuracy (%)
Overall	78.6	84.99	6.39
Open-ended	77.8	83.07	5.27
Close-ended	79.8	82.41	2.61

Qualitative Analysis: Fig. 3 shows the qualitative results of K-PathVQA compared to the VL model. We can observe that the visual features influence the VL model in the image, e.g., the left image (Fig. 3 (a)), although is a gallbladder, share similar visual traits (red color, texture and shape) as a kidney, whereas the middle image (Fig. 3 (b)), appears to be an image of G-cell hyperplasia, and the right image (Fig. 3 (c)) looks

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/jbhi.2023.3294249, IEEE Journal of Biomedical and Health Informatics

similar to myocarditis from its appearance. However, our K-PathVQA model can identify the correct answer by combining the medical knowledge and the language information with the visual features. For example, in the example in Fig. 3 (a), the K-PathVQA model uses the fact that the gallbladder is located in cholecystitis, and the description of cholecystitis from the knowledge used describes that cholecystitis is an inflammatory lesion of the gallbladder to find the correct answers. Similarly, our model used knowledge of that hemosiderosis caused by bleeding (Fig. 3 (b)) and 'rocky mountain spotted' is associated with fever (Fig. 3 (c)), to derive the correct answers.

It is worth noting that the K-PathVQA answers remain consistent from a language semantic perspective, even in the case of incorrect answers. For example, the question in Fig. 4 (a) ''What is distinguished from nodular hyperplasia by its solitary circumscribed nature?", our model predicted 'fibroma' as an answer which means a tumour made up of fibrous tissue . Fibroma is close to the ground truth 'adenoma' because it refers to the tumour associated to a tissue , e.g., fibroma refers to a tumour made up of fibrous tissue and adenoma refers to a tumour of gland tissue. Another example in Fig. 4 (c), "What have not yet formed?", demonstrates that our proposed model answered 'the fibrous tissue' which is close to the ground truth (fibrin nets). Fibrin is defined as "fibrous, non-globular protein involved in the clotting of blood." From these results, we observe that the medical knowledge representation helps capturing of interactions between the image and the question.

V. Discussion

In this study, we proposed a K-PathVQA that incorporates external medical knowledge to improve the performance of question answering in pathology images. We demonstrated the importance of work in medical VQA, as it has the potential to plays a crucial role in medical decision-making, computer-aided diagnosis, and training. The experiments conducted using a publicly available PathVQA dataset showed that K-PathVQA outperformed the best baseline method, achieving significant improvements in accuracy. These results highlight the importance of leveraging external medical knowledge to enhance the performance of PathVQA models.

A limitation of the work is that the model performance is shown to be influenced by the quality of an external medical knowledge graph. This may be a plausible reason for why the performance is relatively low compared to non-medical VQA and suggests future work in improving medical knowledge graphs. A second limitation relates to the computational cost and scalability, which we did not report on here, and warrants further investigation.

VI. Conclusion

This paper presents the K-PathVQA architecture, a novel knowledge-driven approach designed to enhance the performance of PathVQA by incorporating external structured content. K-PathVQA employs medical knowledge to enrich the question representation and then fuses the embeddings resulting from vision, language, and knowledge contents to learn a joint knowledge-image-question representation. Such approach was shown to overcome the lack of domain-specific knowledge in PathVQA models that is instrumental in properly understanding the pathology images. Our method was evaluated on a benchmark dataset of PathVQA, and the experimental results demonstrated its effectiveness over existing state-of-the-art approaches. Future works could focus on enhancing the handling of knowledge noise caused by excessive incorporation of knowledge through the development of improved knowledge distillation techniques. Additionally, incorporating a preprocessing step to exclude healthy tissue and focus on lesions could be explored to further improve accuracy.

References

[1] Muhammad Khalid Khan Niazi, Anil V Parwani, and Metin N Gurcan. Digital pathology and artificial intelligence. *The lancet oncology*, 20(5):e253–e261, 2019.

[2] Jason D Hipp, Anna Fernandez, Carolyn C Compton, and Ulysses J Balis. Why a pathology image should not be considered as a radiology image. *Journal of pathology informatics*, 2, 2011.

[3] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

[4] Xuehai He, Zhuo Cai, Wenlan Wei, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathological visual question answering. *arXiv preprint arXiv:2010.12435*, 2020.

[5] Xuehai He, Zhuo Cai, Wenlan Wei, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Towards visual question answering on pathology images. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 708–718, 2021.

[6] Dhruv Sharma, Sanjay Purushotham, and Chandan K Reddy. Medfusenet: An attention-based multimodal deep learning model for visual question answering in the medical domain. *Scientific Reports*, 11(1):1–18, 2021.

[7] Tuong Do, Binh X Nguyen, Erman Tjiputra, Minh Tran, Quang D Tran, and Anh Nguyen. Multiple meta-model quantifying for medical visual question answering. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 64–74. Springer, 2021.

[8] Usman Naseem, Matloob Khushi, and Jinman Kim. Vision-language transformer for interpretable pathology visual question answering. *IEEE Journal of Biomedical and Health Informatics*, 2022.

[9] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250, 2008.

[10] Kai Zhang, Jun Yu, Zhiling Yan, Yixin Liu, Eashan Adhikarla, Sunyang Fu, Xun Chen, Chen Chen, Yuyin Zhou, Xiang Li, et al. Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *arXiv preprint arXiv:2305.17100*, 2023.

[11] RUFAI ZAKARI, Jim Wilson Owusu, Ke Qin, Hailin Wang, Zaharaddeen Karami Lawal, and Tao He. Vqa and visual reasoning: An overview of approaches, datasets, and future direction. *Datasets, and Future Direction*.

[12] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Qiao Jin, Bhuwan Dhingra, William W. Cohen, and Xinghua Lu. Probing biomedical embeddings from language models, 2019.

[15] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. *arXiv preprint arXiv:1805.07932*, 2018.

[16] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29, 2016.

[17] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016.

[18] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE transactions on neural networks and learning systems*, 29(12):5947–5959, 2018.

[19] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 1821–1830, 2017.

[20] Binh D Nguyen, Thanh-Toan Do, Binh X Nguyen, Tuong Do, Erman Tjiputra, and Quang D Tran. Overcoming data limitation in medical visual question answering. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 522–530. Springer, 2019.

[21] Fuji Ren and Yangyang Zhou. Cgmvqa: a new classification and generative model for medical visual question answering. *IEEE Access*, 8:50626–50636, 2020.

[22] Asma Ben Abacha, Soumya Gayen, Jason J Lau, Sivaramakrishnan Rajaraman, and Dina Demner-Fushman. Nlm at imageclef 2018 visual question answering in the medical domain. In *CLEF (Working Notes)*, pages 1–10, 2018.

[23] Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *CLEF (Working Notes)*, pages 1–11, 2019.

[24] Lei Shi, Feifan Liu, and Max P Rosen. Deep multimodal learning for medical visual question answering. In *CLEF (Working Notes)*, 2019.

[25] Xin Yan, Lin Li, Chulin Xie, Jun Xiao, and Lin Gu. Zhejiang university at imageclef 2019 visual question answering in the medical domain. In *CLEF (Working Notes)*, pages 1–9, 2019.

[26] Yalei Peng, Feifan Liu, and Max P Rosen. Umass at imageclef medical visual question answering (med-vqa) 2018 task. In *CLEF (Working Notes)*, pages 1–9, 2018.

[27] Yangyang Zhou, Xin Kang, and Fuji Ren. Employing inception-resnet-v2 and bi-lstm for medical domain visual question answering. In *CLEF (Working Notes)*, pages 1–11, 2018.

[28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[29] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[30] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[31] Zhibin Liao, Qi Wu, Chunhua Shen, Anton van den Hengel, and Johan Verjans. Aiml at vqa-med 2020: Knowledge inference via a skeleton-based sentence mapping approach for medical domain visual question answering. pages 1–14. CLEF, 2020.

[32] Li-Ming Zhan, Bo Liu, Lu Fan, Jiaxin Chen, and Xiao-Ming Wu. Medical visual question answering via conditional reasoning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2345–2354, 2020.

[33] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[34] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[35] Usman Naseem, Imran Razzak, Shah Khalid Khan, and Mukesh Prasad. A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models. *arXiv preprint arXiv:2010.15036*, 2020.

[36] Usman Naseem, Adam G Dunn, Matloob Khushi, and Jinman Kim. Benchmarking for biomedical natural language processing tasks with a domain specific albert. *arXiv preprint arXiv:2107.04374*, 2021.

[37] Usman Naseem, Byoung Chan Lee, Matloob Khushi, Jinman Kim, and Adam G Dunn. Benchmarking for public health surveillance tasks on social media with a domain-specific pretrained language model. *arXiv preprint arXiv:2204.04521*, 2022.

[38] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining, 2019.

[39] Usman Naseem, Matloob Khushi, Vinay Reddy, Sakthivel Rajendran, Imran Razzak, and Jinman Kim. Bioalbert: A simple and effective pre-trained language model for biomedical named entity recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2021.

[40] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations, 2018.

[41] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE, 2021.

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, and booktitle=Advances in neural information processing systems pages=5998–6008 year=2017 Kaiser, Lukasz and Polosukhin, Illia. Attention is all you need.

[43] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[44] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[45] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

VII. Appendix A

TABLE X

Comparison of different pre-trained CNNs and LMs used for question and image representation.

Overall				
Image\Question	BioELMo	BioBERT	BLUEBERT	BERT
ResNet50	**68.97**	65.78	63.91	60.90
Inception	64.79	61.70	60.67	58.36
DenseNet	61.91	59.54	58.93	57.78
VGG19	60.64	58.64	56.52	56.64
Open-ended				
Image\Question	BioELMo	BioBERT	BLUEBERT	BERT
ResNet50	**42.12**	41.13	40.20	39.52
Inception	37.87	36.86	35.63	34.59
DenseNet	36.83	35.59	34.38	33.98
VGG19	35.89	34.48	33.70	32.70
Closed-ended				
Image\Question	BioELMo	BioBERT	BLUEBERT	BERT
ResNet50	**94.60**	90.10	87.76	86.43
Inception	93.14	89.56	87.42	85.64
DenseNet	92.86	88.94	85.51	84.03
VGG19	91.82	87.85	85.63	83.91