Vision-Language Transformer for Interpretable Pathology Visual Question Answering

Usman Naseem, *Student Member, IEEE*, Matloob Khushi, and Jinman Kim, *Member, IEEE*

***Abstract*— Pathology visual question answering (PathVQA) attempts to answer a medical question posed by pathology images. Despite its great potential in healthcare, the technology is still in its early stages. It is not widely adopted because it requires both high and low-level interactions on both the image (vision) and question (language) to generate an answer. Existing methods focused on treating vision and language features independently, which were unable to capture these high and low-level interactions. Further, these methods failed to offer capabilities to interpret the retrieved answers, which are obscure to humans. Despite this, models' interpretability to justify the retrieved answers has remained largely unexplored. Interpretability has become important to engender users' trust in the retrieved answer by providing insight into the model prediction. Motivated by these limitations, this paper introduces a vision-language transformer that embeds vision (images) and language (questions) features for an interpretable PathVQA. We present an interpretable transformer-based Path-VQA (TraP-VQA), where we embed transformers' encoder layers with vision and language features extracted using pre-trained CNN and domain-specific language model (LM), respectively. A decoder layer is then embedded to upsample the encoded features for the final prediction for PathVQA. Our experiments showed that our TraP-VQA outperformed the state-of-the-art comparative methods with the public PathVQA dataset. Furthermore, our additional experiments validated the robustness of our model on another medical VQA dataset, and the ablation study demonstrated the capability of our integrated transformer-based vision-language model for PathVQA and the robustness of our model on another medical VQA dataset. Finally, we conclude by discussing the interpretability of each component of Trap-VQA by presenting the visualization results of both text and images, which explains the reason for a retrieved answer in the PathVQA.**

***Index Terms*— Pathology Images, Interpretability, Visual Question Answering, Vision-Language**

I. Introduction

PATHOLOGY examines the causes and effects of diseases or injuries and involves diagnosing conditions through specimens surgically removed from the body, such as organs, tissues, bodily fluids, and in some cases, the whole body[1]. These specimens are routinely captured using images, as exemplified in Fig. 1. Pathology visual question answering (PathVQA) can answer many questions, including necrosis, inflammation, and cancer diagnosis. Despite rapid advances in the use of machine learning for medical image analysis, e.g., such as X-rays, MRIs, and CT scans, with results often surpassing or on-par with clinicians, there has been a paucity of research for pathology images [1]. This is partly due to the complexity in combining the pathology imaging data with the accompanying clinical question and the answer.

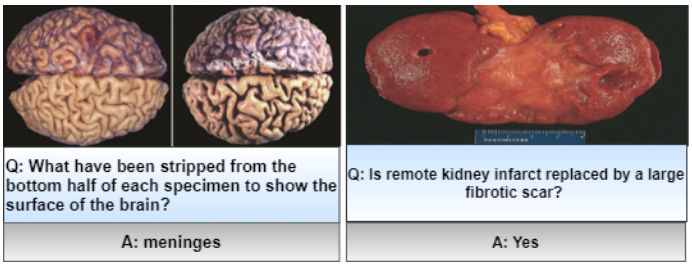


Fig. 1. Examples of a PathVQA where questions are asked in two types: (right) closed-ended questions where the answers are restricted to "Yes" or "No" and, (left) open-ended questions where the answers are in free-form text such as keyword.

In VQA, given an image accompanied by a relevant question, an algorithm deals with both the image (vision) and the question (language) and predicts an answer to the question (see Fig. 1). VQA is a challenging task as it demands an in-depth understanding and high-level interactions on both the input image and the question of both structural language and non-structural image to generate the answer. Because of this vision-language (VL) analysis property, which is commonly used in medical applications, e.g., decision support systems and for medical training, VQA is generating growing interest in the medical imaging community. Interpretability is crucial to producing convincing answers for the system's reliability and efficiency of medical visual question-answer (MedVQA).

To mitigate the limitations of modeling long dependencies and support parallel processing of sequence in recurrent neural networks (RNNs), Vaswani et al. [2] proposed an encoder-decoder based architecture built on multi-head self-attention and feed-forward neural networks referred to as transformer for machine translation tasks. It is the state-of-the-art (SOTA) method in various natural language processing

Authors would like to acknowledge contribution to this research from the Australian Research Council (ARC) grants.

U. Naseem, M. Khushi, J. Kim are with the School of Computer Science, The University of Sydney, Australia. MK is with University of Suffolk, Ipswich, UK (e-mail: {usman.naseem,matloob.khushi, jinman.kim}@sydney.edu.au).

[1]https://www.mcgill.ca/pathology/about/definition

(NLP) tasks [3] and models such as Bidirectional Encoder Representations from Transformers (BERT) [4], Generative Pre-trained Transformer (GPT) [5] and others [6]–[8] have built upon this architecture. Inspired by the success of NLP and the strong representation capabilities of transformers, it has attracted tremendous interest and proved its dominance over convolutional networks (CNNs) [9] in computer vision (CV) [3], including medical imaging, to leverage the self-attention mechanisms to fuse information across the whole image, considering the image's low-level features. However, there is a paucity of research that combines both the encoder-decoder layers of the transformer in 'vision-language' tasks, such as for MedVQA, which can complement the benefit of leveraging the transformer's architecture.

Existing approaches [10]–[18], though improved the accuracy but failed to capture high and low-level interactions from both image and text that are important to retrieve a correct answer. Furthermore, these methods did not leverage both encoder-decoder layers of a transformer to fuse both image and text features for a MedVQA task. These methods did not provide users with appropriate interpretations of the retrieved answer. To provide interpretability, it is important to capture the interaction of both image and text features and understand question-answer pairs. To develop an interpretable PathVQA model, we address the following challenges: (i) capture both high and low-level representation of question-answer pair, and (ii) interpretability is as important as accuracy for MedVQA, which is unfortunately often neglected.

In this paper, we present a novel and interpretable vision-language transformer, which leverages both the encoder and the decoder layers of a transformer to embed vision and language features for the MedVQA (PathVQA in our case) task. We propose TraP-VQA – a transformer-based pathology visual question answering method where we embed low-level image features with domain-specific contextual information derived from the questions, which are then used to answer the question. TraP-VQA uses the strength of CNNs to extract image features at low-level, domain-specific language model (LM) to extract domain-specific contextual information, and transformer to capture global dependencies at high-level. Extensive experiments show the advantage of our model against other methods and establish the new SOTA results on a popular benchmark PathVQA dataset.

The remainder of the paper is structured as follows: A summary of the related work is provided in section II. Section III presents the methodology of the proposed model. Experiments details and the results are then presented in section IV. Finally, section V concludes the study.

II. Related Work

A. Medical Visual Question Answering (MedVQA)

Existing works in MedVQA [10]–[14], [19], especially methods used in the ImageCLEF challenges [11], [20], [21] on MedVQA tend to adapt the advance methods used in general-domain VQA such as Multi-modal Compact Bilinear (MCB) [22], Stacked Attention Networks (SAN) [23], Bilinear Attention Networks (BAN) [24], Multi-modal factorized bilinear (MFB) [25] and Multimodal Factorized High-order (MFH) [26]. Typically, pre-trained models such as ResNet [27] or VGGNet [28] are used to extract image features, recurrent neural networks (RNNs), such as long short-term memory (LSTM) [29] and gated recurrent unit (GRU) [30], word embeddings, and Bidirectional Encoder Representations from Transformers (BERT) [4] are adopted for extracting text-based features. In first edition of ImageCLEF challenge[2], Peng et al. [13] used ResNet-152 and LSTM for extracting image and text-based features, respectively and adopted MFH for LV features concatenation. Zhou et al. [14] adopted Inception-Resnet-v2 and BiLSTM to model features from both image and text, respectively, and fused the encoded questions with the image features to predict the answers. Abacha et al. [11] employed pre-trained VGG-16 and LSTM for extracting image and text features and later used SAN to combine the question and image features. In the second edition of ImageCLEF challenge[3], the best model [12] adopted BERT and pre-trained VGG-16 for text and image features, respectively, and used MFB for fusing the VL features. In the third edition of ImageCLEF challenge[4], the best method [15] detected the question type by dividing questions into two types, i.e., open-ended and close-ended type and transformed the VQA task into a simplifier multi-task image classification problem.

However, approaches tested on general-domain VQA for MedVQA undergo data scarcity and lack of multilevel reasoning ability due to discrepancies between medical and general domains. To overcome the issue of limited data, Nguyen et al. [16] presented the Mixture of Enhanced Visual Features (MEVF) component, which utilizes the Model-Agnostic Meta-Learning (MAML) [31] and the Convolutional Denoising Auto-Encoder (CDAE) [32] to solve the data limitation by initializing the model weights for image feature extraction. To enable VQA models to learn reasoning skills due to the disparity of questions, Zhan et al. [17] proposed Question-Conditioned Reasoning (QCR) and Type Conditioned Reasoning (TCR) modules and applied the MEVF image backbone. Most recently, to adapt the model to a different form of output, Ren and Zhou proposed CGMVQA [18] - a novel classification and generative model for MedVQA which integrated both a classifier and a generator and adopted the multi-head self-attention method of a transformer. The CGMVQA model was tested on VQA-Med-2019 and outperformed the winner of the VQA-Med-2019 challenge. However, these methods are not interpretable. In contrast to the previous studies, our approach is designed to generate vision and language interpretations in the context of PathVQA. To the best of our knowledge, it is the first work that fuses both encoder and decoder layers to fuse vision and language features and provide interpretation for retrieved answers.

B. Transformers in Vision-Language Tasks

Recently, researchers from the VL community have also adopted transformers [2], e.g., for video captioning, visual

[2]https://www.imageclef.org/2018/VQA-Med
[3]https://www.imageclef.org/2019/medical/vqa/
[4]https://www.imageclef.org/2020/medical/vqa

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: DOI 10.1109/JBHI.2022.3163751, IEEE Journal of Biomedical and Health Informatics

commonsense reasoning (VCR), and VQA. Some of the examples include Vision, and Language BERT (ViLBERT) [33], and Learning Cross Modality Encoder Representations from Transformers (LXMERT) [34]; both of these models used two-stream BERT with a VL co-attention component to model the vision and language inputs. ViLBERT is fine-tuned on various downstream tasks, including image-to-text retrieval, referring expression, and VQA, whereas a pre-trained model of LXMERT, is fine-tuned on only Visual Reasoning for Real (NLVR) and VQA. Other related works such as VisualBERT [35] and VL-BERT [36] used a single stream of a transformer to model visual and image-text relation for tasks like VQA and VCR. Due to BERTs' tremendous success and popularity, researchers focused only on the encoder part of the transformer in all of the studies mentioned above, leaving the decoder layer unutilized. Conversely, we present a unified method using both the encoder and decoder layers of a transformer and fully leverage the benefit of a complete transformer architecture.

III. Method

Problem Definition: First, we define our problem formally, given a pathology image I and a relevant question Q; the goal of the PathVQA task is to predict the answer $\hat{A}$. Mathematically, it can be formulated as:

$$\hat{A} = f(I, Q, \theta) \tag{1}$$

where $theta$ denotes the model parameters, and f is the answer prediction function.

Overview of the Architecture: TraP-VQA consists of four main components as shown in Fig. 2: (1) question feature extraction using domain-specific LM (BioELMO) with BiLSTM to capture contextual information; (2) image feature extraction using ResNet with CNN to capture low-level features; (3) Transformer Encoder used to fuse the extracted image (vision) and question (language) features, and to model high-level global features and, (4) Transformer Decoder to upsample the encoded features for final prediction.

A. Question Feature Extraction

BioELMO with BiLSTM: BioELMo [37] is trained on 10M PubMed biomedical abstracts text and has the same network structure as ELMo.

ELMo is a task-specific concatenation of these features learned from Bi-LM, where all layers are flattened to a single vector (equation 2).

$$ELMo_n^{\text{task}} = E(M_n; \Theta^{\text{task}}) = \gamma^{\text{task}} \sum_{j=0}^{L} s_j^{\text{task}} h_{h,j}^{LM} \tag{2}$$

where s^{task} are softmax normalised weights for concatenation of several layer representations and γ^{task} is a hyper parameter for representation optimisation and scaling.

We used pre-trained BioELMo to extract the contextual features of the given questions Q, as given by equation. (3). BioELMo largely outperforms ELMo and previous SOTA methods in a variety of biomedical text mining tasks.

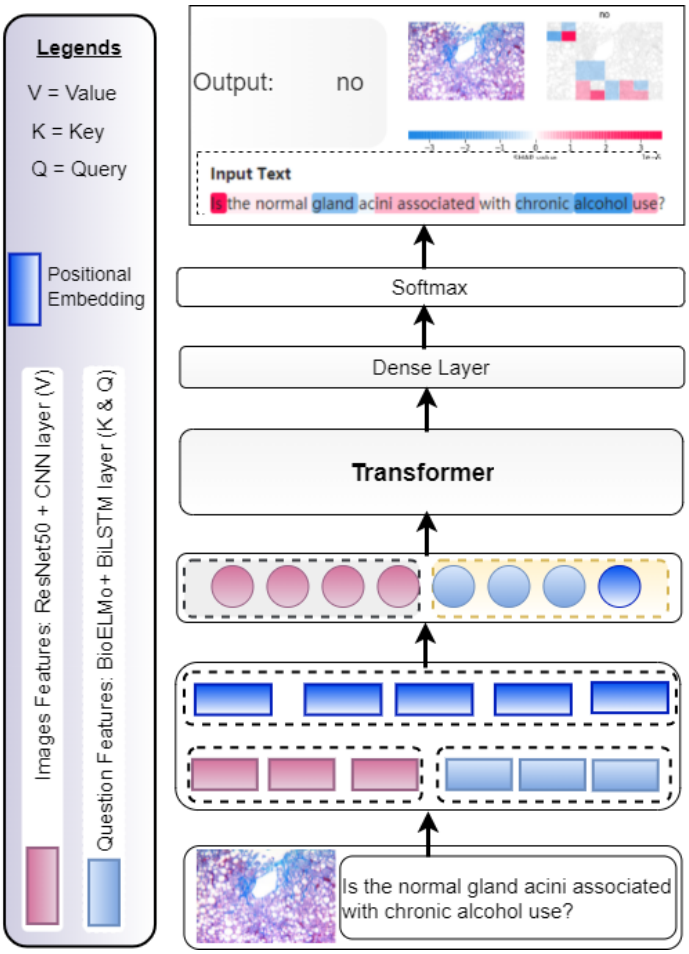


Fig. 2. Overall architecture of interpretable TraP-VQA.

$$X_Q = BioELMo(Q) \tag{3}$$

Using BioELMo, a 1,024 dimensional vector X_Q, is fed to a BiLSTM layer to model the information from both directions and outputs a hidden representation h_i at a given time i, by concatenating the hidden representations from both forward $\overrightarrow{h_i}$ and backward $\overleftarrow{h_i}$ LSTM (equation 4).

$$h_i = [\overrightarrow{h_i} \parallel \overleftarrow{h_i}] \tag{4}$$

where $\parallel$ indicate the concatenate operator, $X_Q \in \mathrm{R}^{l*d}$, l is the question length, d is the vector size for each word. X_Q is padded to match the maximum questions length l_{max} to X_Q^{pad} before forwarding to BiLSTM layer (equation 5).

$$X_Q^l = BiLSTM(X_Q^{pad}); \tag{5}$$

where, $X_Q^{pad} \in \mathcal{R}^{lmax*d}$, $X_Q^l \in \mathcal{R}^{lmax*512}$; X_Q^l will go through a denser layer, a positional encoding, and a dropout layer and outputs a matrix of question features represented by X_Q^f.

where $X_Q^f \in \mathrm{R}^{lmax*512}$. To extract the question features, we experimented with various general and domain-specific LM

such as ELMo, BERT, BioBERT, and BLUEBERT. BioELMO performed best (see Table III) as compared to others.

B. Image Feature Extraction

ResNet with CNN: We extracted image features using pre-trained VGG19, InceptNet, DenseNet, and ResNets and identified that pre-trained ResNet50 performed best compared to others (see Table III). We reshaped an image I to match the shape of ResNet50 (224, 224, 3). Since we did not need ResNet50 to act as a classifier but rather as a feature extractor, we dropped the last three fully connected layers, retaining only the output of the last average pooling layer as image features X_I (equation 6).

$$X_I = ResNet50(I) \tag{6}$$

X_I is fed to another 2-D CNN layer of kernel size 3, the activation function of ReLu, and forwarded to a dense layer to shrink the channel, reshaping and flattening is to maintain as much information as possible and outputs a matrix of image features represented by X_I^l, as given by equation (7). This structure retains as much information as possible while matching the first dimension of the image feature matrix X_I^l to the first dimension of the question feature matrix X_Q^f.

$$X_I^l = Convolution2D(ResNet50(X_I); \tag{7}$$

where; $X_I^l \in \mathcal{R}^{7*7*512}, X_I^l \in \mathcal{R}^{lmax*512}$

C. Transformer

Transformer comprises of an encoder(s)-decoder(s) structure. Each encoder layer is comprised of a multi-head self-attention and a feed-forward neural network. Like the encoder, the decoder has three sublayers, two of which are similar to the encoder (multi-head self-attention and feed-forward), while the third sublayer carries out multi-head attention on the encoder's outputs. The input vector is first transformed into three different vectors: the value vector *v*, the key vector *k* and the query vector *q*. Vectors derived from different inputs are then combined together into 3 different matrices, namely, *V*, *K*, and *Q* (equation 8).

$$Att(Q, K, V) = softmax(\frac{Q.K^T}{\sqrt{d_k}}).V \tag{8}$$

where $Q \in \mathcal{R}^{Lxd}$, $K \in \mathcal{R}^{Lxd}$d, $V \in \mathcal{R}^{Lxd}$, L is the sequence length and d is the feature depth.

The query is passed to the component, which searches for the most similar key and returns the value related to that key. Two matrix multiplications, as well as a softmax function, help to speed up the process. softmax$(Q.K^T)$ generates a probability distribution with peaks at locations that are positioned by the keys for the relating query. This serves as a pseudo-mask, and by matrix multiplying it with V, we can get the centralized values that the network should pay attention to in the first place.

1) Transformer Encoder: We used only the transformer encoder layer to fuse the image (vision) and question (language) features extracted in previous steps. In the original transformer encoder, the input (V, Q, K) to the encoder is a sequence of words that we modified and replaced with image and question features.

At the first encoder layer, the image feature matrix X_I^l, is used as V, and the question feature matrix X_Q^f is forwarded as an input of both Q and K with positional encoding. At the second encoder layer, we again used X_I^l as input V and the output of the first encoder layer is forwarded to Q and V of the second encoder. Here the input V, Q, and V are processed in the same way as in the original transformer encoder layer.

2) Transformer Decoder: We used the same transformer decoder layer as the original transformer to upsample the encoded features. Here, again we used two decoder layers for the final prediction.

At first, a one-hot vector of $< start >$ token will go through a trainable embedding layer, and positional encoding is fed to the decoder layers. The softmax function will give a probability distribution of each one-hot vector. The decoder will take the one-hot vector with the highest probability and append the corresponding vocabulary to the answer. The decoding process continues until the decoder generates the one-hot vector of the $< end >$ token. Here the working mechanism of decoder layers is the same as in the original transformer decoder layers.

IV. Experiments and Results

A. Datasets

We experimented with data sets from PathVQA [38]. The question design was inspired by the examination of the American Board of Pathology (ABP), with the aim to capture medical questions that are a part of a certified pathologist testing in the US. The PathVQA contains 4,998 images and 32,799 QA pairs. The images with captions are extracted from textbooks and online digital libraries. The question are divided into seven categories: what (40.9%), where (4.0%), when (0.9%), whose (0.6%), how(3.0%), how much/how many (0.9%), and, yes/no (49.8%). The first six categories' questions are open-ended, including 16,465 in total and accounting for 50.2% of all questions. The last category is the close-ended (yes or no) question. We used the standard training, validation and test set provided in [38] to evaluate our model.

B. Experimental Settings

We trained our model using Adam [39] optimizer with a learning rate of 0.0001 and with a batch size of 64 for 20 epochs. We used the grid-search optimization technique to tune the optimal parameters. We tested with different transformer layers (see Table **??** in appendix) and used accuracy as an evaluation metric.

C. Baseline Methods

We compared our results with the following baselines.

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: DOI 10.1109/JBHI.2022.3163751, IEEE Journal of Biomedical and Health Informatics

General VQA Methods:
BAN [24]: embeds image and textual features using a Gated Recurrent Unit and a Faster R-CNN network and learns bilinear attention distributions employing bilinear attention networks (BAN) and approximates the bilinear interaction between question and image embeddings using low-rank approximations.
MCB [22]: A CNN encodes the image, while an LSTM network encodes the questions and answers. An attention mechanism is used to infer the answer using a multimodal compact bilinear (MCB) pooling method.
SAN [23]: With CNN and LSTM, the stacked attention network (SAN) locates image regions that are useful to answer the question. It queries the image several times to narrow the region to be observed.
MFB [25]: embeds images and questions/answers using CNN and LSTM and uses Multi-Modal Factorized Bilinear (MFB) pooling to fuse question and image features.
MEVF [16]: extracts image and language features using CNN and LSTM and uses a mixture of enhanced visual features (MEVF) with SAN and BAN to fuse visual and language features.
Other Methods:
Vision language: We used SOTA vision language-based models such as LXMERT [34], VisualBERT [35] and UniTER [40] to fuse image and the language features extracted using CNN and LSTM.
CMSSL [41]: a SOTA approach for PathVQ, which detects and ignores noisy self-supervised examples from pretraining to learn robust visual and textual features.

D. Results

Table I presents the results of TraP-VQA and the baselines TraP-VQA achieved the best performance and outperformed all the baselines. Compared to the following closed method (UniTER), our method achieves 64.82% accuracy, an absolute increase of 4.49% for the overall task and 37.72% accuracy for open-ended question type, which is an absolute increase of 2.39% compared to second-best (LXMERT). Furthermore, 93.57% accuracy is an absolute increase of 5.87% compared to the second-best method (UniTER) for closed-ended question types. We observed that general VQA methods such as MFB, SAN, MCB, and BAN perform poorly compared to transformer-based methods. Although MEVF+BAN and MEVF+SAN perform better than the base BAN and SAN methods, performance is less than that of transformer-based methods.

Our results showed that our TraP-VQA consistently outperformed all the baselines. These results demonstrate the effectiveness of the transformers in capturing global relationships, as evident from the baselines failing to capture them; the attention maps generated by the scaled dot-product attention module in a transformer highlight the image region responsible for each generated text token. This performance improvement can be thought to be due to the presented framework's ability to encode low-level visual features using a convolutional neural network (ResNet50) and a domain-specific language model (BioELMo) for text representation trained on relevant corpora, as well as its ability to leverage the powerful transformer capability in modeling the global relationship.

TABLE I
COMPARISON OF PROPOSED METHOD V/S THE BASELINES.

Model	Overall	Open-ended	Close-ended
MFB [25]	39.85	20.15	53.77
SAN [23]	42.43	23.40	59.40
MCB [22]	57.04	29.03	57.60
BAN [24]	55.10	33.50	68.20
MEVF +SAN [16]	57.10	25.87	86.90
MEVF +BAN [16]	57.90	26.75	87.50
LXMERT [34]	60.00	35.33	83.00
VisualBERT [35]	60.08	33.03	86.99
UniTER [40]	60.33	33.79	87.70
LXMERT+CMSSL [41]	60.10	34.50	87.10
BAN+CMSSL [41]	58.40	33.50	87.20
Ours	**64.82**	**37.72**	**93.57**

E. Ablation Study

1) Effects of using text only features with transformer: Following [41], we fused text only features with a transformer to analyze the impact on the performance from the absence of the images. Table II shows the results of using text-only features extracted from different LMs with a transformer. We observed a 10.72% drop in performance (from 64.82% to 54.10%) in the overall task. For open-ended question types, a 16.44% drop in performance was observed (from 37.72% to 21.28%), whereas, in closed-ended question types, a minor drop (1.92%) is observed when we used BioELMO to extract text features. Performance dropped in all other cases when we used text-only features (ranging from 9.98% to 11.25%) overall task, 16.44% to 19.70% in open-ended question types, and 0.72% to 3.65% in closed-ended question types. Although the accuracy drops when using only the text features obtained by BioELMo, the performance is better than the overall accuracy of 2 baselines (MFB and SAN) when compared to using the full model (TraP-VQA). We attribute these due to the fact that most questions do not require visual content to answer questions. This drop-in accuracy shows the importance of using both VL features in our model.

2) Effect of different VL models with transformer: We replaced different pre-trained VL models to extract features and fused them to the transformer layer. Table III shows the results from fusing the transformer with different models to extract the image and question features. The optimal combination was found to be the use of ResNet50 and BioELMo in all tasks. The performance varies from 51.79% to 64.82%, 9.57% to

TABLE II
COMPARISON OF FUSING ONLY LANGUAGE FEATURES (TEXT ONLY) WITH TRANSFORMER.

Model	Overall	Open-ended	Close-ended
BioELMo	**54.10**	**21.28**	**91.65**
BioBERT	47.53	11.83	78.38
BLUEBERT	46.79	11.05	74.74
BERT	46.92	12.22	76.78

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: DOI 10.1109/JBHI.2022.3163751,

TABLE III
COMPARISON OF FUSING DIFFERENT PRE-TRAINED CNNS AND LMS USED FOR VL FEATURE EXTRACTION WITH TRANSFORMER.

Image\Question	Overall				Open-ended				Close-ended			
	BioELMo	BioBERT	BLUEBERT	BERT	BioELMo	BioBERT	BLUEBERT	BERT	BioELMo	BioBERT	BLUEBERT	BERT
ResNet50	**64.82**	58.78	56.91	56.90	**37.72**	31.53	30.24	30.72	**93.57**	79.10	81.76	80.43
Inception	51.79	49.03	47.67	48.36	9.57	12.06	12.63	10.59	92.14	83.56	81.42	82.64
DenseNet	57.71	54.54	52.93	51.78	23.63	21.59	21.38	20.98	93.26	80.94	82.51	82.03
VGG19	58.64	56.74	56.52	55.64	33.69	26.48	27.70	27.70	92.82	79.85	80.63	77.91

37.72% and 92.14% to 0 93.57% in accuracy for overall, open and closed-ended tasks respectively when BioELMO is fused with different pre-trained CNN models to a transformer. We observed that in all cases, ResNet50's performance was more prominent compared to others. In addition, it had the highest consistency among the different LMs. Further, for question features, BioELMo outperformed all other methods in extracting question features compared to SOTA BERT and biomedical versions of BERT on all tasks. This is expected, as BioELMo is proven to be a better-fixed feature extractor and also better at clustering similar information than BERT-based models for extraction question features [37].

F. Robustness to other MedVQA

To evaluate the robustness of our method, we designed experiments using other MedVQA datasets. We note that there was no other public pathology VQA dataset in our thorough investigation. Hence, we used SLAKE [42], a MedVQA dataset of radiology images. This dataset is different because it ensures the diversity of modalities (e.g., CT, MRI, and X-Ray), covered body parts (e.g., head, neck, and chest), and question types (e.g., vision-only, knowledge-based, and bilingual). SLAKE is a comprehensive MedVQA dataset with semantic labels and a structured medical knowledge base annotated by expert physicians (see Fig. **??** in an appendix for examples). Fig. **??** in the appendix presents the detailed results, showing that our method outperformed the baselines, including the SOTA method used in [42] on other MedVQA datasets due to its ability to capture global relationships between image-question pairs. These results validate the robustness of our method on MedVQA tasks.

G. TraP-VQA Interpretation

Using existing state-of-the-art interpretable tools such as Gradient-weighted Class Activation Mapping (Grad-CAM) [43] and SHapley Additive exPlanations (SHAP) [44], we perform a qualitative evaluation for visual, textual, and transformer attention interpretations on the PathVQA dataset. In addition, we illustrate qualitative examples of multi-modal interpretations.

1) Qualitative Evaluation for Textual Interpretation: To evaluate the use of BioELMO compared to other LMs, we performed interpretable qualitative analysis. Fig. 3 illustrates the visual representation of different textual features extracted using Word2Vec, BioELMo, BERT, BioBERT, and BLUEBERT used in our model. We used K-means [45] to cluster the textual embedding into the two-dimensional feature space. It is clear to observe from Fig. 3 that BioELMo embeddings show separable distributions compared to other LMs (Fig. 3). This demonstration visually shows the efficacy of using BioELMo embeddings in our model. In Table II, we further show the importance of textual features in our model and demonstrated that BioELMo performed better compared to other language features.

2) Qualitative Evaluation for transformers' attention layer Interpretation: To quantify the proposed transformer as a fusion layer, we performed an interpretable qualitative analysis. In Fig. 4, we show a pathology image (left) with a relevant question (top) and SHAP values at the bottom of the image. High visual scores (attention weights) are shown in blue, whereas low visual scores are shown in red. The visual scores are taken as values from attention weights from the last decoder layer of the transformer. We observe that in closed-ended question types, TraP-VQA assigned high visual scores to words like 'pemphigus' and 'gland.' These words are directly associated with their pathology image counterpart; for example, top left shows the image of dilation and hypertrophy in pemphigus vulgaris and gland acini for the bottom left image. Similarly, in the open-ended question types, the visualization shows that our model gives more weight to words like 'Where' (top right), 'What,' and 'present' (bottom right) according to our intuition and low visual scores determiners and prepositions. This visualization explains the reason behind the retrieved answer and how TraP-VQA assigns more weight to important words relevant to pathology images.

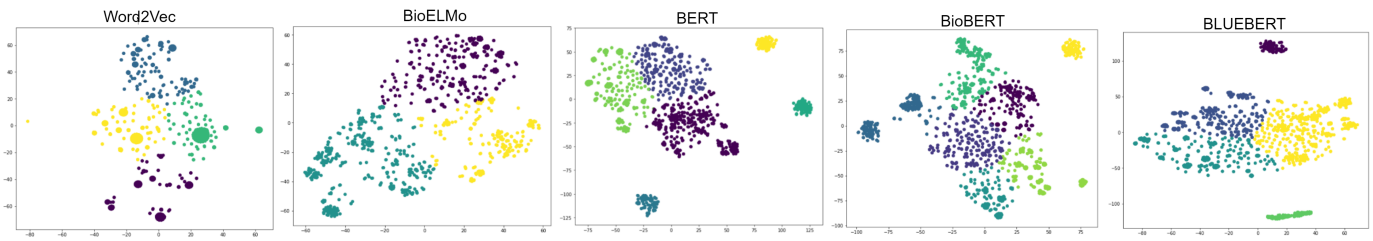


Fig. 3. Interpretation (Visualization) of textual features obtained using different language models.

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: DOI 10.1109/JBHI.2022.3163751, IEEE Journal of Biomedical and Health Informatics

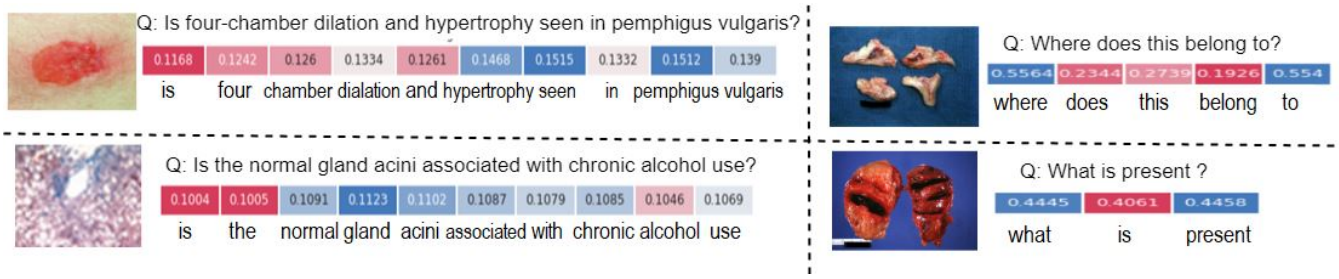
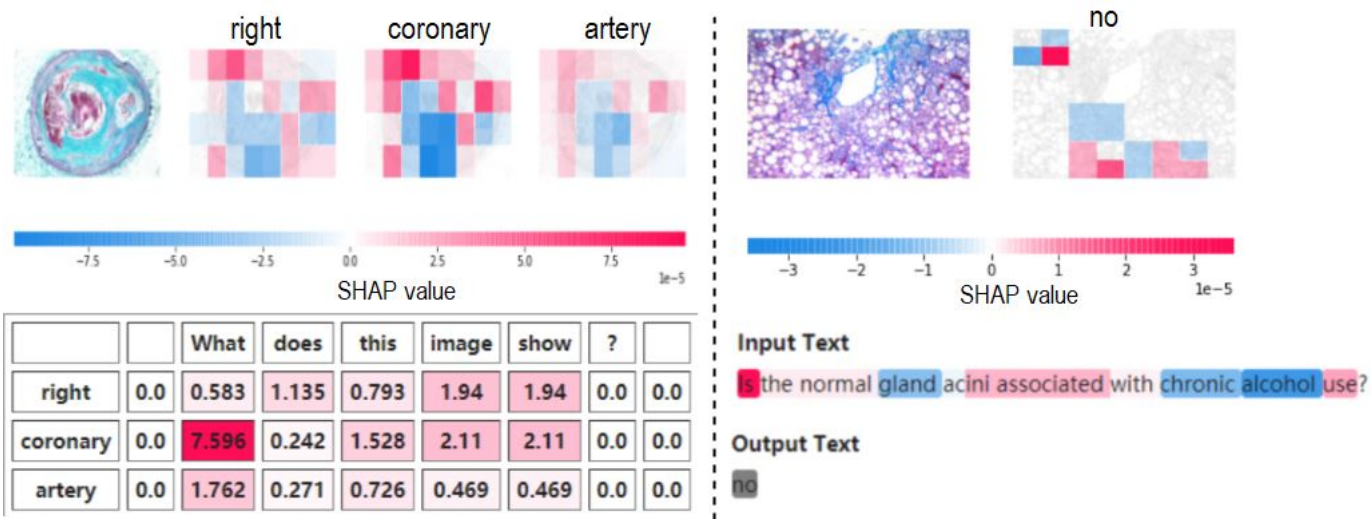


Fig. 4. Visual scores of words in transformers attention. We demonstrate the attention weight of the last transformer layer as the raw visual scores. Blue (high) and red (low) represent high and low visual scores.



		What	does	this	image	show	?	
right	0.0	0.583	1.135	0.793	1.94	1.94	0.0	0.0
coronary	0.0	7.596	0.242	1.528	2.11	2.11	0.0	0.0
artery	0.0	1.762	0.271	0.726	0.469	0.469	0.0	0.0

Fig. 5. Trap-VQA Visualization using SHAP. The red color represents high scores, whereas blue shows low scores.

3) Qualitative Evaluation for Trap-VQA Interpretation: In Fig. 5 and Fig. 6, we show outputs of visual and textual interpretations for the unseen classes on PathVQA. Fig. 5 is an example of visual interpretations obtained using SHAP. SHAP is a game-theoretic method for explaining any machine learning model's output. It uses the traditional Shapley values from game theory and associated extensions to connect optimal credit allocation with local explanations and final output visualization. In Fig. 5, we present a pathology image at the top with SHAP values and a relevant question at the bottom, also with SHAP values. High visual scores are indicated by red, whereas low visual scores are shown by blue. From the visual scores at the bottom of the pathology images in Fig. 5 (left), we see that TraP-VQA gives more weight (red) to relevant words such as 'coronary' for the question *'What does this image show?*' and 'right coronary artery' region is appropriately highlighted in red pixels (boxes) in the image (top left). However, in the close-ended question (*'Is the normal gland acini associated with chronic alcohol use?*') shown in Fig. 5 (right), a high visual score is assigned to the question word 'Is' and predicts a correct answer 'no .'This interpretation of TraP-VQA is aligned with the attention weight visualization shown in Fig. 4 and explains the reason for the retrieved answer.

To evaluate the use of ResNet in TraP-VQA compared to other pre-trained CNNs, we performed an interpretable qualitative analysis. Fig. 6 visualizes the extracted image features (column 2 to column 5) using Grad-Cam. We can see that different models emphasized different parts of the image, and ResNet (column 2) best associated the region of interest when compared to other CNN-based models. We attribute this to ResNet having a deeper (50 layers) model compared to InceptNet, DenseNet, and VGG19. This is consistent with ResNet having the best feature extraction performance as in Table III.

Furthermore, Fig. 6 presents the randomly selected examples where TraP-VQA correctly predicted the answers. A pathology image and a relevant question are shown in the left (column 1), and the visualization of different CNN models and predicted answering using different models are shown in the right columns 2 to 5. TraP-VQA focuses on the region of interest corresponding to the attributed label, whereas other models failed in some predictions. Correct answers are indicated in green, while incorrect answers are shown in red.

V. Conclusion

This paper presented a TraP-VQA method that embeds the image and question features, coupled with domain-specific contextual information, via a transformer for PathVQA. We used Grad-Cam and SHAP to interpret our retrieved answers visually to indicate which area of the image contributed to the predicted answer. We show that using ResNet in our

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: DOI 10.1109/JBHI.2022.3163751,

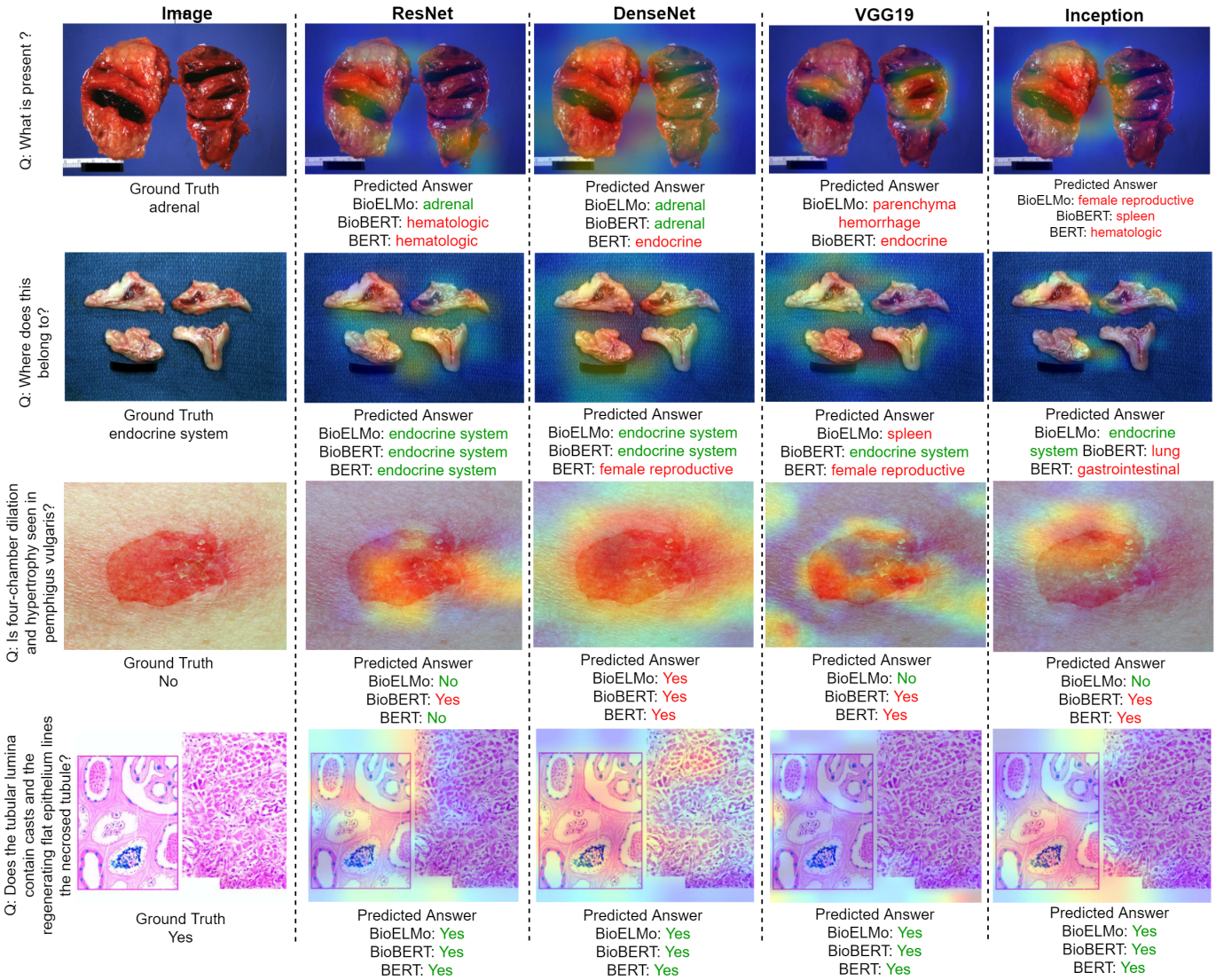


Fig. 6. Qualitative results of TraP-VQA: Original pathology images and their associated relevant questions are shown in the left column 1, whereas, visualization of different CNN models using Grad-Cam and predicted answering using different models are shown in the right column 2 to column 5. Correct answers are indicated in green, while incorrect answers are shown in red

model focuses on the region of interest. In contrast, other models sometimes focus on the wrong part of the image. For textual interpretations, we visually show that text embeddings obtained using domain-specific language models have clear separable distributions compared to other language models tested. In addition, visualization of the transformers' attention showed proposed model assigns more weight to the relevant words and explains the reason for the retrieved answer. Empirical evaluation of the popular benchmark dataset of PathVQA demonstrated that our method achieved superior performance relative to state-of-the-art comparative models and ensured adequate evidence to interpret the retrieved answers.

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] 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.

[3] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on visual transformer. *arXiv preprint arXiv:2012.12556*, 2020.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[5] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2018.

[6] 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. *Transactions on Asian and Low-Resource Language Information Processing*, 20(5):1–35, 2021.

[7] 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.

[8] 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.

[9] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[10] 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)*, 2018.

[11] 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)*, 2019.

[12] 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)*, 2019.

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

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

[15] 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. CLEF, 2020.

[16] 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.

[17] 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.

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

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

[20] Sadid A Hasan, Yuan Ling, Oladimeji Farri, Joey Liu, Henning Müller, and Matthew P Lungren. Overview of imageclef 2018 medical domain visual question answering task. In *CLEF (Working Notes)*, 2018.

[21] Asma Ben Abacha, Vivek V Datla, Sadid A Hasan, Dina Demner-Fushman, and Henning Müller. Overview of the vqa-med task at imageclef 2020: Visual question answering and generation in the medical domain. *CLEF 2020 Working Notes*, pages 22–25, 2020.

[22] 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.

[23] 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.

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

[25] 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.

[26] 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.

[27] 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.

[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] Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555, 2014.

[31] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017.

[32] Jonathan Masci, Ueli Meier, Dan Cireşan, and Jürgen Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. In *International conference on artificial neural networks*, pages 52–59. Springer, 2011.

[33] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019.

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

[35] 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.

[36] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

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

[38] 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.

[39] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[40] 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.

[41] 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.

[42] 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.

[43] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[44] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[45] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.