Leveraging LIME for Transparent and Accurate Visual Question Answering in Education

Anuradha Yeole
Dept. of Computer Science and Engineering (Data Science)
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India
anuradhasudhiryeole@gmail.com

Krish Valecha
Dept. of Computer Science and Engineering (Data Science)
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India
saumyakrish123@gmail.com

Jai Vasi
Dept. of Computer Science and Engineering (Data Science)
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India
jai.n.vasi1108@gmail.com

Kanchan Dabre
Dept. of Computer Science and Engineering (Data Science)
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India
kanchandabre@gmail.com

***Abstract*—With the advancement of Artificial Intelligence (AI), several sectors including computer vision and natural language processing, among others, have seen significant progress. One of the hot areas in AI research is Visual Question Answering (VQA) where a question posed is answered by taking a closer analysis of both visual and textual information. In spite of VQA systems working reasonably effectively through incorporating Convolutional Neural Networks (CNNs) to address the visual elements of the task then followed up with Long Short-Term Memory (LSTM) networks to analyse the text understanding, these frameworks still have problems related to how explainable and intelligible the outcomes are. Interaction with complex AI technology has its limitations especially in the field of high-impact applications like education, healthcare, legal and others where there is need to explain to humans the reasons for various conclusions reached. In this research work, the authors provide an innovative approach through a vision Transformer (ViT) based VQA model integrated with LIME, with an intention of addressing the two issues inherent in traditional AI approaches i.e. accuracy and explainability. LIME which facilitates the understanding of the model is used in the context of the ViT model to address the question, what aspects of visual and textual inputs were most significant for the model. By incorporating Local Interpretable Model-Agnostic Explanations (LIME), accuracy increased from 62.5% to 68.2% and Macro F1 from 60.3% to 66.7%. Wu-Palmer Similarity (WUPS) also improved, with WUPS@0.9 rising from 63.8% to 69.0% and WUPS@0.0 from 70.5% to 75.1%. These results demonstrate that LIME enhances both model performance and interpretability, paving the way for more reliable AI systems in real-world, high-risk settings.**

***Index Terms*—Visual Question Answering (VQA), Explainable AI (XAI), Local Interpretable Model-Agnostic Explanations (LIME), Vision Transformers**

I. Introduction

Visual Question Answering (VQA) has emerged as a key research area in AI, combining Computer Vision (CV), Natural Language Processing (NLP), and Knowledge Representation (KR). It enables AI systems to process both visual and textual inputs, offering a unified framework to interpret and respond to questions about images [1, 2, 15, 16].

Early VQA models primarily relied on Convolutional Neural Networks (CNNs) for image feature extraction and Long Short-Term Memory (LSTM) networks for question encoding [1, 2]. While promising, these models faced difficulties with long-range dependencies and complex scene comprehension.

Recent advancements include Vision Transformers (ViTs), which process images as sequences of patches, allowing for better handling of long-range dependencies through multi-head attention mechanisms [3, 4, 5]. These models have shown improved performance in scenarios requiring complex scene understanding.

Similarly, the shift from LSTMs to transformer-based encoders in NLP has enhanced question comprehension [6, 7]. Transformers use self-attention to capture dependencies within text, enabling the handling of complex and ambiguous queries. Multimodal transformers such as ViLBERT further integrate visual and textual data more effectively, improving accuracy in VQA tasks [7].

The integration of visual and textual information remains a major challenge in VQA. Recent models utilize complex attention mechanisms to dynamically merge visual concepts with related language descriptions [4, 7, 15]. This has deepened the understanding of visual-linguistic interactions in VQA systems.

As VQA research continues to evolve, interpretable models have gained importance. Techniques like LIME and visual attention mechanisms provide human-understandable explanations of predictions, critical in sensitive areas such as healthcare [8, 9, 10].

Adaptive learning and feedback mechanisms, often based on reinforcement learning, have also been explored to allow VQA models to improve over time based on user feedback [11, 12]. This leads to more robust and flexible systems capable of learning from real-world inputs.

Domain-specific VQA models, particularly in medical imaging and remote sensing, have shown enhanced performance by leveraging domain knowledge [13, 14]. These models highlight the benefits of tailoring VQA systems to specific applications.

Emerging trends include addressing multilingual VQA challenges [16], incorporating visual common sense for broader AI tasks [17], and developing more complex neural architectures [18]. These advancements indicate a shift towards VQA systems with greater versatility and wider AI applicability.

In summary, VQA research continues to make significant strides, driven by innovations in image and language processing, multimodal learning, and interpretability. These developments point to a future where VQA systems become integral to AI applications, advancing human-AI interaction.

II. VQA Dataset Description And Collection

This research utilizes a dataset designed for Visual Question Answering (VQA) tasks, comprising 14,457 images paired with educational questions from domains such as natural science, social science, and language science. The dataset is organized into 9,998 folders, each containing images linked to specific questions. For example, questions about physical states are accompanied by images depicting different phases of matter, while geography-related queries feature maps or symbolic illustrations. This structure emphasizes the need for multimodal models to effectively integrate visual and textual information, as the questions range from straightforward factual inquiries to complex reasoning tasks.

The dataset features multiple-choice questions with 3 to 5 answer options, where only one is correct. Each answer has been manually validated, ensuring that the corresponding image is essential for arriving at the correct response. The questions are categorized into three subject areas: natural science, covering basic scientific concepts; social science, addressing geography and social studies; and language science, focusing on grammar and comprehension. Additionally, images are stored in uniquely named folders, with each folder containing an average of 2-3 relevant images, supporting diverse training data while providing unseen validation and test data to evaluate model generalization. Future iterations of the dataset may explore various answer types beyond multiple-choice, expanding its applicability in educational contexts.

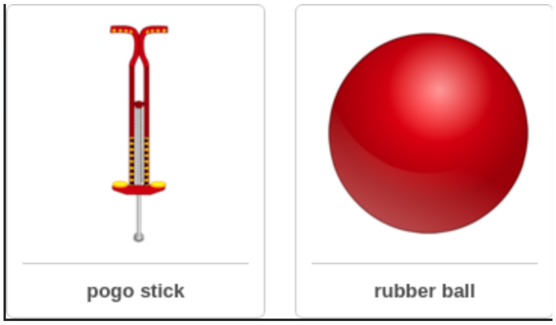


Fig. 1: Question: What object in the image is used for bouncing?
Choices: Pogo stick, Rubber ball, Both, None
Answer: Pogo stick.
Source: Image taken from the dataset used in study.

III. Gaps in existing works

The field of Visual Question Answering (VQA) has made significant strides in recent years, largely due to the integration of advanced machine learning models like Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). These models have enhanced the ability of AI systems to interpret images and answer questions based on the visual content. Despite these advancements, there remains a critical gap in terms of model transparency and interpretability, especially in high-stakes environments like education. While accuracy has been a major focus of VQA research, the ”black-box” nature of most AI models continues to pose a substantial challenge. In applications such as education, where decisions made by AI systems can have far-reaching consequences, it is not enough for the model to simply provide an accurate answer—it must also offer clear, understandable reasons for its predictions. This lack of transparency in AI models creates a barrier to trust, making it difficult for users, such as educators or students, to accept AI-driven answers without understanding how those answers were derived.

Most of the existing VQA systems fail to integrate explainable AI (XAI) techniques that allow users to interpret the reasoning behind the model’s predictions. While there have been notable efforts to apply attention mechanisms in VQA models to visualize the areas of an image that contributed to a decision, these explanations often fall short in providing meaningful insights. Attention maps, for instance, may highlight certain regions of an image but do not explain why those specific areas are relevant to the question being asked. Furthermore, many of the current VQA models, including those using CNNs and ViTs, often treat the interpretability aspect as an afterthought, focusing primarily on optimizing accuracy. As a result, these models leave much to be desired in terms of providing clear, understandable explanations that could help users trust and validate the system’s outputs.

Additionally, existing work on VQA primarily focuses on general-purpose datasets that are not tailored to specific domains. While models like ViLBERT have achieved high accuracy in diverse VQA tasks, they are not designed with particular educational goals in mind. For instance, questions in education often require more than just factual knowledge—they may involve reasoning, conceptual understanding, and explanations that are crucial for learning. In this context, VQA models should not only be able to answer questions correctly but also provide meaningful explanations that align with human reasoning. However, the majority of existing models lack this capability, making them less suited for educational applications where the goal is not just to provide an answer but to help students understand the rationale behind it.

Furthermore, while there has been progress in integrating reinforcement learning in various machine learning tasks, very few VQA models incorporate adaptive learning based on user feedback. Without such mechanisms, these systems are unable to improve their performance over time based on real-world interactions, making them less robust in dynamic environments. Feedback from users, especially in the educational domain, can be crucial for refining AI models and ensuring they align more closely with human expectations and understanding. The integration of feedback loops in VQA systems is an area that has been largely unexplored, leaving a gap in the research that could significantly improve the usability and adaptability of VQA systems.

In contrast to these existing limitations, this research presents a novel approach by combining Vision Transformers with Local Interpretable Model-Agnostic Explanations (LIME). By integrating LIME into a ViT-based VQA system, we address both the challenges of accuracy and transparency simultaneously. LIME allows for clear, region-specific explanations that explain why certain visual elements in the image influenced the model's decision, making it possible for users to understand the model's reasoning process. This capability is especially crucial in educational applications, where the ability to explain the reasoning behind AI-driven decisions is paramount. The introduction of LIME in this study represents a significant step forward in the field of VQA, bridging the gap between high-performing models and transparent, trustworthy AI systems that can be effectively deployed in real-world educational contexts.

This work also addresses the gap in domain-specific VQA applications. The model is trained on a dataset designed specifically for educational purposes, incorporating images and questions from natural science, social science, and language science. This domain-specific approach allows the model to perform better in educational environments, where the questions are more structured and tied to curriculum-based learning. By designing a VQA model tailored to educational contexts, we ensure that the AI can handle questions that require not only factual knowledge but also reasoning and explanation. This approach has the potential to transform how AI is used in educational settings, providing students and educators with a more reliable, transparent, and effective tool for learning and teaching.

IV. Methodology

Here, a structured VQA system containing both vision and text input, where the former is processed by an encoder using a Vision Transformer (ViT) and the latter is processed by a text encoder for the question, and the final answer is presented. A key component of the system is the integration of a feedback mechanism using Local Interpretable Model-Agnostic Explanations (LIME) to improve model interpretability and adjust the prediction based on user feedback. This section outlines the complete process, starting from the input image and question, to the final output and feedback loop.

The model, as shown in Figure 2, processes input images using a Vision Transformer (ViT) to generate visual features, while text is encoded separately. A multi-head attention mechanism combines both modalities, and LIME is applied to generate interpretable visual explanations for the model's predictions.

A. Input Image Processing

The visual component in a VQA system is essential for establishing the context for questions. In this model, input images, drawn from diverse educational contexts (e.g., medical, academic, general knowledge), are preprocessed for compatibility with the Vision Transformer (ViT).

Preprocessing includes rescaling images to 224x224 pixels and normalizing based on the following mean and standard deviation values:

$$\text{Mean: } [0.485, 0.456, 0.406]$$

$$\text{Standard Deviation: } [0.229, 0.224, 0.225]$$

This ensures uniformity for feature extraction by the ViT. Data augmentation techniques, including segmentation, further enhance image training for accurate visual element capture needed to answer the questions.

1) Patch Generation

After preprocessing, the next step for ViT is patch generation. Unlike CNNs, ViTs divide images into smaller patches and process them sequentially, similar to how language models handle sequences. Each image is divided into fixed-size patches (e.g., 16x16 pixels), which act as tokens for the transformer model. These patches are flattened and embedded into higher-dimensional space, akin to word embeddings in NLP. This allows the ViT model to capture long-range dependencies between patches, improving its ability to understand images for tasks like visual question answering.

B. Vision Transformer (ViT) Encoder

The Vision Transformer (ViT) processes image patches generated in the previous stage by applying self-attention mechanisms. ViTs enable both global and local relationships to be captured, unlike CNNs that focus primarily on local features. This capability allows the system to answer complex queries by considering a broader context. The ViT encoder, therefore,

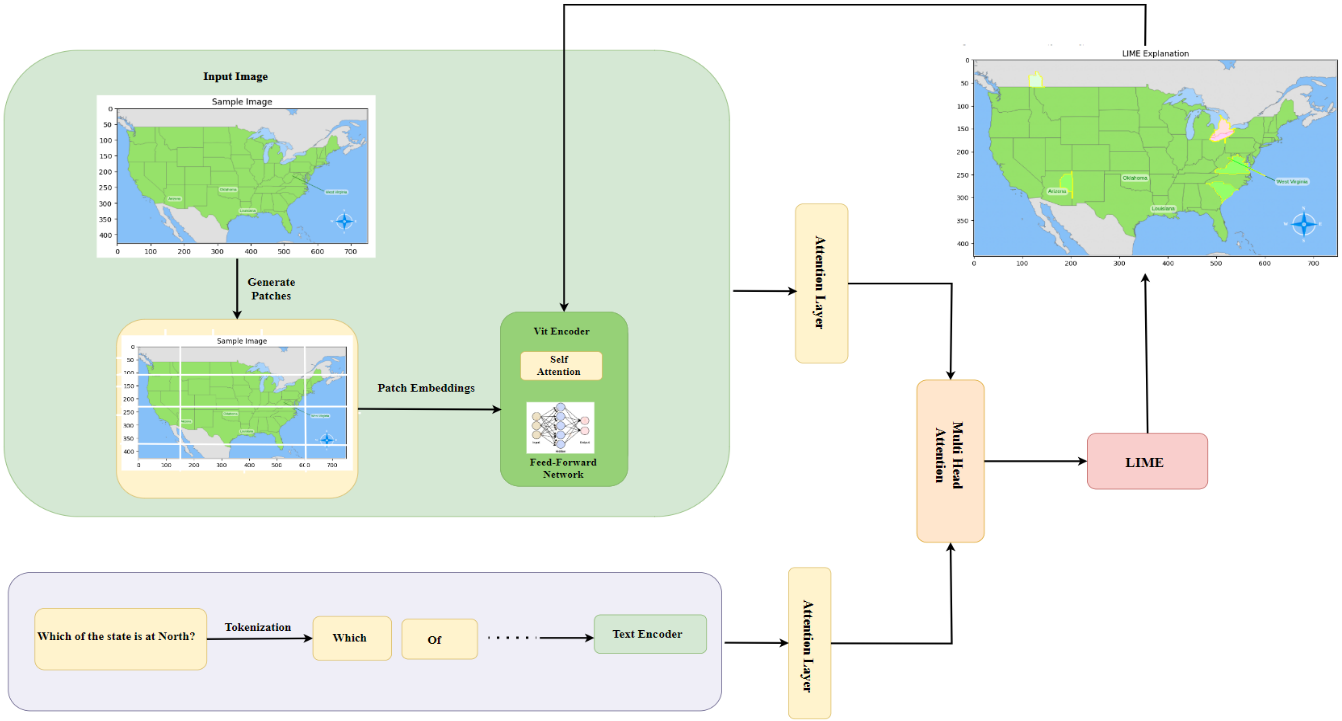


Fig. 2: System architecture of proposed model

plays a crucial role in accurately interpreting visual input, improving reasoning ability and answer accuracy.

In a ViT, the input image I of size $H \times W \times C$ (height, width, and number of channels) is divided into fixed-size patches of $P \times P$. Each patch is flattened into a 1D vector and projected into a higher-dimensional space using a linear projection. The total number of patches N is given by:

$$N = \frac{H \times W}{P^2} \tag{1}$$

Each patch is embedded into a vector of size D (the hidden size of the transformer), forming a patch embedding matrix of size $N \times D$. Positional embeddings are added to retain each patch's relative position as transformers are permutation-invariant models.

1) Self-Attention Mechanism

Self-attention forms the core of transformer architectures, enabling the model to focus on important image regions relevant to answering the question. Each patch attends to every other patch, capturing long-range dependencies and improving overall image understanding. The attention mechanism follows these steps:

1) **Query, Key, and Value Matrices:** Patch embeddings are projected into Query (Q), Key (K), and Value (V) matrices using linear transformations:

$$Q = X \cdot W_Q, \quad K = X \cdot W_K, \quad V = X \cdot W_V \tag{2}$$

where X is the patch embeddings and W_Q, W_K, W_V are weight matrices.

2) **Scaled Dot-Product Attention:** Attention between patches is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) \tag{3}$$

where d_k is the dimensionality of the keys. The softmax ensures attention weights sum to 1, allowing the model to focus on relevant patches.

2) Multi-Head Attention

To enhance the model's ability to capture different relationships across the image, multi-head attention is employed. The input is divided into h separate attention heads, each with its own set of learned weights:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h) \cdot W_O \tag{4}$$

Each head computes attention scores independently, and their outputs are concatenated and linearly transformed by W_O. This mechanism allows the model to attend to multiple parts of the image simultaneously, improving its comprehension of the overall image context.

3) Feed-Forward Network

After multi-head attention, the output passes through a feed-forward network (FFN) to refine the learned features. The FFN consists of two fully connected layers with ReLU activation in between, which helps filter out irrelevant details and enhance key visual features. The transformation applied to each patch embedding is:

$$\text{FFN}(x) = \text{ReLU}(x \cdot W_1 + b_1) \cdot W_2 + b_2 \quad (5)$$

Here, W_1, W_2, b_1, and b_2 are learned parameters. This ensures that the final patch representations are more meaningful and relevant to the input question.

C. Text Processing

The other major aspect in the VQA system would be the text processing pipeline. The output of the text processing pipeline will align the visual features that are extracted by the ViT to what the user is asking. This means that the model is reading the question properly, which actually allows an appropriate association between the image and the answer.

1) Input Question Tokenization

Tokenization breaks down the user's question into smaller units (words or subwords) for the text encoder. This step converts the question into a sequence of tokens, enabling the model to process and interpret the structure and meaning of the question. Proper tokenization is crucial as it helps the model align the text input with the visual features extracted from the image.

2) Text Encoder

After tokenization, the question is converted into a dense, high-dimensional vector representation using the text encoder. This transformation captures the semantics of the question, including entities, their relationships, and overall intent. The text encoder is based on transformer layers, similar to the ViT encoder, ensuring consistent architecture and facilitating alignment between image and text representations in subsequent stages.

D. Attention Mechanism

Now comes the attention mechanism to merge these two visual and textual representations. It aligns both modalities, so that model will focus on the right parts of the image based on the question.

1) Attention Layer

The attention layer integrates information from both the image and the question to determine the most relevant answer. Outputs from the ViT encoder (visual representation) and the text encoder (textual representation) are fed into this layer. It assigns weights to different parts of the image based on their relevance to the question, ensuring that critical image patches align with key components of the question. The attention layer computes attention scores for the question tokens and image patches as follows:

$$\text{CrossAttention}(Q, K, V) = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) \cdot V \quad (6)$$

Here, the queries (Q) are derived from the question representation, while the keys (K) and values (V) come from the image patches.

2) Multi-Head Attention

In order to capture different aspects of the image-question pair, several heads are used in parallel. Here the attention layer uses multi-head attention; that is, several attention mechanisms are applied in parallel to capture various aspects of the image and question. Every head attends to parts of the image and question differently so that the model captures a rich set of interactions between the two modalities. This would result in having a more complete knowledge of the concerned pair image-question, thus allowing the model to give right answers.

E. Explanation via LIME

This justification of model predictions will help the users generate trust, especially in educational and decision-making setups. LIME is a crucial constituent of this VQA system, as it imparts explainability to predictions generated from the model. A deep-learning model like this one, especially one like VQA, suffers from what might be called the 'black-box' problem. This architecture presents LIME to get over this problem through offering human-understandable explanations of the model's decisions.Below shows an example image for a random query from the dataset, and Figure 3 shows the LIME explanation for that image.

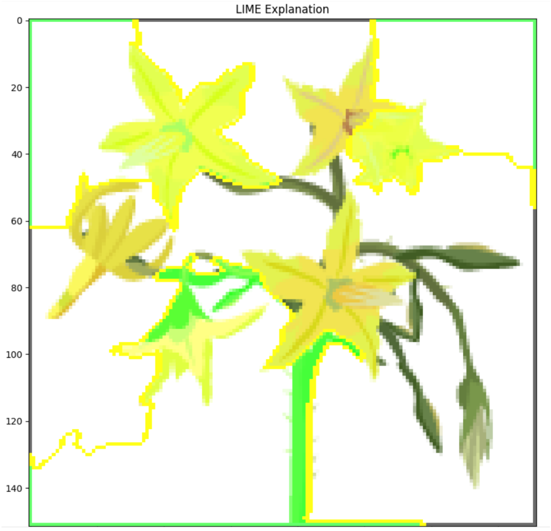


Fig. 3: LIME explanation generated by the VQA model for the query.

1) Generation of LIME Explanation

Once the model produces a prediction, LIME is used to explain the decision. That is the step that makes the model's decisions interpretable to users. Once the multi-head attention layer produced a prediction, LIME generates an explanation for the output by analyzing the input image and question in a local context. It determines which parts of the image and which tokens in the question contributed most to the final answer. This step allows users to understand the reasoning behind a model's choice of a particular answer and thus inducts the VQA system an even more transparently interpretable system.

LIME explanations also become an important feedback mechanism for improvement in the performance of the model.

F. Feedback Mechanism

Feedback mechanisms facilitate continuous learning and improvement of the VQA model. User input regarding the correctness and satisfaction of the model's answers plays a crucial role in fine-tuning. When an answer is deemed incorrect or unsatisfactory, this feedback is relayed back to the ViT encoder. The information is utilized to adjust weights within the attention layers and encoders, enabling the model to learn from its mistakes through this feedback loop. Consequently, with ongoing training, the model becomes more accurate and robust.

Additionally, the incorporation of LIME (Local Interpretable Model-agnostic Explanations) enhances the feedback process by providing understandable and interpretable insights into the model's predictions. LIME highlights the most relevant parts of the image that contributed to a given answer, allowing the feedback mechanism to focus on regions that may have been misestimated. This identification of critical image areas leading to incorrect answers helps the model learn to respond better to similar questions in the future.

Through feedback coupled with LIME explanations, the model corrects erroneous answers not only quantitatively but also qualitatively, aligning with user expectations in reasoning. This refinement process results in increasingly valid and trustworthy outcomes as the model gains experience.

Algorithm 1 Pseudocode for VQA Model Training with LIME Explanations

```
model ← initialize_vqa_model()
dataset ← load_and_preprocess_data()
tokenizer, feature_extractor ← load_nlp_components()
repeat
    batch ← get_next_batch(dataset)
    text_features ← tokenizer(batch.text)
    image_features ← feature_extractor(batch.images)
    lime_scores ← generate_lime_explanations(batch)
    logits ← model(text_features, image_features, lime_scores)
    loss ← compute_loss(logits, batch.labels)
    gradients ← backpropagate(loss)
    update_model_weights(model, gradients)
    if validation_step then
        metrics ← evaluate_model(model, validation_data)
        if $metrics > best\_metrics$ then
            save_model(model)
            best_metrics ← metrics
        end if
    end if
until convergence or max_epochs reached
final_model ← load_best_model()
test_results ← evaluate_model(final_model, test_data)
return test_results
```

V. RESULTS

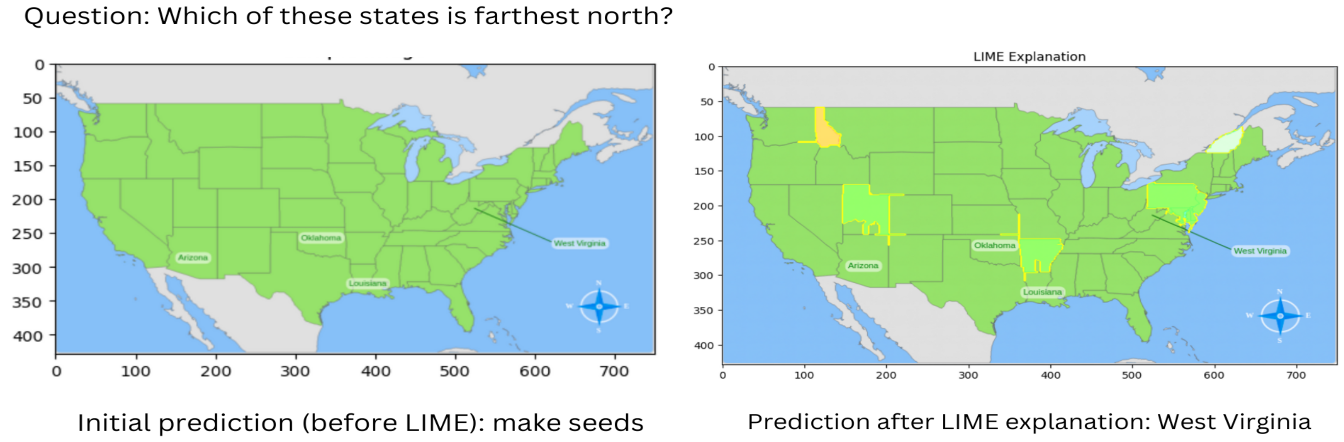
This section highlights the effectiveness of the Vision Transformer-based VQA model, both before and after LIME has been applied. The performance comparison has been made in terms of LIME's influence on the ability of the model to produce more accurate and interpretable predictions. This happens through multiple performance metrics; hence, the level of improvement from the addition of explainability in the VQA system is obtained. LIME doesn't just provide qualitative explanations for the predicted output of a single instance but also enables the model to concentrate on the most relevant features in a complex visual input, hence providing answers that are more credible and contextually correct.

In order to evaluate the feasibility of the proposed model, we apprise its performance with the currently outstanding available VQA models. Earlier studies exercising convolution neural nets and LSTMs focusing on tasks of visual question answering (VQA) were performed with some success but often on the price of clarity. Most of these cases, include the ViLBERT models, which brings together vision and language through the use of transformer architecture which leads to great performance in the VQA tasks, have great accuracy level in VQA tasks; however, such models are still unable to provide clear, interpretable grounds whenever predictions are made due to the evidence provided in the models. This is especially common for educational pedagogues who only require straightforward explanations for the decisions made.

On the other hand, the ViT based VQA model with LIME offered better predictions while advancing the need for visual explanations in further improving transparency. In this case, LIME contributed towards increasing the model's accuracy from 62.5% to 68.2%. Furthermore, the meaning that model would be able to provide for its decisions; when confidence was provided through AI done models such as WUPS offers a better chance of supporting educational based applications. Understanding of the images by the model and questions have shifted more closely towards perception of humans.

Evaluation Metrics

The effectiveness of the Vision Transformer-based VQA model is assessed through three main evaluation metrics: **Accuracy**, **Macro F1 Score**, and **Wu-Palmer Similarity (WUPS)**. These metrics provide a comprehensive view of the model's predictive capabilities before and after the inclusion of LIME. As shown in Table 2, significant improvements were observed across all metrics post-LIME integration.


Question: Which of these states is farthest north?

Initial prediction (before LIME): make seeds

Prediction after LIME explanation: West Virginia

Fig. 4: Predictions before and after LIME explanation

TABLE I: Evaluation metrics before and after applying LIME for the Vision Transformer-based VQA model

Evaluation Metric	**Before LIME**	**After LIME**
Accuracy	62.5%	**68.2**%
Macro F1 Score	60.3%	**66.7**%
WUPS@0.9	63.8%	**69.0**%
WUPS@0.0	70.5%	**75.1**%

Visual Question Answering Results with LIME

In Figure 4, demonstrates the predictions made by the Vision Transformer-based VQA model before and after integrating LIME for explainability. The question presented to the model was, *"Which of these states is farthest north?"* The initial prediction made by the model, before applying LIME, was "make seeds," indicating that the model struggled to focus on the relevant regions of the image and made an irrelevant prediction due to the lack of interpretability.

After applying LIME, the model highlighted the relevant regions of the map, focusing on the correct areas that corresponded to the question. This improved focus led to the correct prediction of *"West Virginia"* as the state farthest north. The LIME explanation, visualized through highlighted regions, demonstrates how the model adjusted its attention to the appropriate states in response to the question.

VI. Conclusion

A Vision Transformer-based Visual Question Answering system was proposed. It was designed to overcome the challenges of answering open-ended questions about images. Given an image, combined with a question in natural language, the problem consisted in generating the correct answer that matches the content visual and textual of the image. The model was here trained on a dataset with diversified sets of images and questions, mainly from domains such as education, general knowledge, and science.

The incorporation of LIME allowed the explanation of the model using terms that are more humanly understandable, which was important in the application of this model in realms like education or health. Since real-world application of AI models may have severe potential consequences, increasing the transparency of LIME improves usability in high-risk applications.

Future Scope

The results demonstrated that incorporating LIME into the VQA system led to a notable increase in key performance metrics. Accuracy improved from 62.5% to 68.2%, while the Macro F1 score rose from 60.3% to 66.7%. Additionally, Wu-Palmer Similarity (WUPS) scores also increased, with WUPS@0.9 improving from 63.8% to 69.0%, and WUPS@0.0 rising from 70.5% to 75.1%. These improvements illustrate the value of LIME in guiding the model's attention to critical visual elements, thereby improving the overall quality of the predictions.

This work suggests several promising avenues for further research. The model can be made applicable to a wider audience by incorporating a greater variety of educational content in subsequent datasets, such as multiple languages. As it stands, the model relies on a single-language educational dataset but incorporating multilingual capabilities would significantly expand its effectiveness in areas with considerable variation in language use.

Additionally, SHAP (SHapley Additive exPlanations) and Grad-CAM methods can also be used in future iterations of the model to interate other techniques for visual explanations as they provide alternate modalities of visual explainations. Such a comparison would be useful in ascertaining the most suitable methods for providing accurate and useful educational explanations.

Furthermore, it is also crucial to deploy the model in real-world settings and gather user feedback in order to understand the practical strengths and weaknesses of this VQA system in real educational contexts. Such studies can help point out areas of concern such as model bias or difficulties arising from the need to explain certain kinds of complex or overly vague questions.

References

[1] S. Antol et al., ”VQA: Visual Question Answering,” 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 2425-2433, doi: 10.1109/ICCV.2015.279.

[2] Dosovitskiy, Alexey et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.” ArXiv abs/2010.11929 (2020): n. pag.

[3] M. Malinowski, M. Rohrbach and M. Fritz, ”Ask Your Neurons: A Neural-Based Approach to Answering Questions about Images,” 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 1-9, doi: 10.1109/ICCV.2015.9.

[4] Radford, Alec et al. “Learning Transferable Visual Models From Natural Language Supervision.” International Conference on Machine Learning (2021).

[5] Lu, Jiasen Batra, Dhruv Parikh, Devi Lee, Stefan. (2019). ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. 10.48550/arXiv.1908.02265.

[6] Ribeiro, Marco Singh, Sameer Guestrin, Carlos. (2016). “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. 97-101. 10.18653/v1/N16-3020.

[7] Xu, Kelvin Ba, Jimmy Kiros, Ryan Cho, Kyunghyun Courville, Aaron Salakhutdinov, Ruslan Zemel, Richard Bengio, Y.. (2015). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention.

[8] W. Samek, et al., ”Interpretable AI: Current Issues and Future Directions,” arXiv preprint arXiv:1904.13000, 2019.

[9] Y. Li, et al., ”Reinforcement Learning for User Feedback in Visual Question Answering,IEEE Transactions on Neural Networks and Learning Systems, 2019.

[10] P. Gao, et al., ”Dynamic Fusion With Intra- and Inter-Modality Attention Flow for Visual Question Answering,” in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019 pp. 6632-6641. doi: 10.1109/CVPR.2019.00680

[11] P. Anderson et al., ”Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering,” 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 6077-6086, doi: 10.1109/CVPR.2018.00636.

[12] Yu, Zhou et al. “Beyond Bilinear: Generalized Multimodal Factorized High-Order Pooling for Visual Question Answering.” IEEE Transactions on Neural Networks and Learning Systems 29 (2017): 5947-5959.

[13] S. Banik, et al., ”Domain-Specific Visual Question Answering for Medical Imaging,” Proceedings of the 26th ACM International Conference on Multimedia, 2018.

[14] Lobry, S., Marcos, D., Murray, J., Tuia, D. (2020). RSVQA: Visual Question Answering for Remote Sensing Data. IEEE Transactions on Geoscience and Remote Sensing, 58(12), 8555-8566. https://doi.org/10.1109/TGRS.2020.2988782

[15] Z. Yang, et al., ”Stacked Attention Networks for Image Question Answering,” IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[16] Gao, Haoyuan Mao, Junhua Zhou, Jie Huang, Zhiheng Wang, Lei Xu, Wei. (2015). Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question Answering.

[17] X. Lin and D. Parikh. Don’t Just Listen, Use Your Imagination: Leveraging Visual Common Sense for Non-Visual Tasks. In CVPR, 2015.

[18] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In ICCV, 2015.