Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

EDUVQA – Visual Question Answering: An Educational Perspective

Dipali Koshti[1,*], Ashutosh Gupta[1], Mukesh Kalla[1], Pramit Kanjilal[2], Sushant Shanbhag[2], Nirmit Karkera[2]

[1] Department of Computer Science and Engineering, Sir Padampat Singhania University, Udaipur, Rajasthan, India
[2] Department of Electronics and Computer Science, Fr. Conceicao Rodrigues College of Engineering, Mumbai, India

ARTICLE INFO

Article history:
Received 5 July 2023
Received in revised form 1 December 2023
Accepted 20 February 2024
Available online 26 March 2024

Keywords:

Educational VQA; Fact-based VQA; Domain-specific VQA

ABSTRACT

Increasing applications of artificial intelligence in the field of education have changed the way school children learn various concepts. Educational Visual Question Answering or EDUVQA is one such application that allows students to interact directly with images, ask educational questions, and get the correct answer. Two major challenges faced by educational VQA are the lack of availability of domain-specific datasets and often it requires referring to the external knowledge bases to answer open-domain questions. We propose a novel EDUVQA model developed especially for educational purposes and introduce our own EDUVQA dataset. The dataset consists of four categories of images - animals, plants, fruits, and vegetables. The majority of the currently used techniques focus on the extraction of picture and question characteristics in order to discover the joint feature embeddings via multimodal fusion or attention mechanisms. We propose a different method that aims to better utilize the semantic knowledge present in images. Our approach entails building an EDUVQA dataset using educational images, where each data point is made up of an image, a question that corresponds to it, a valid response, and a fact that supports it. The fact is created in the form of <S,V,O> triplet where 's' denotes a subject, 'v' a verb, and 'o' an object. First, an SVO detector model is trained on EDUVQA dataset capable of predicting the Subject, Verb, and Object present in the image-question pair. Using this <S,V,O> triplet, the most relevant facts from our fact base are extracted. The final answer is predicted using these extracted facts, image, and question attributes. The image features are extricated using pretrained ResNet and question features using a pre-trained BERT model. We have optimized and improved on the current methodologies that use a Relation-based approach and built our SVO-detector model that outperforms current models by 10%.

1. Introduction

Visual Question answering is a task where a machine is expected to generate a correct answer to the question based on any image [1]. The examples of Visual question answering has been depicted in Figure 1. It is quite easy for humans to give answers to the posed question. However, for a machine, it needs to extract various features like textual features and image features and perform the joint

[*] *Corresponding author.*
E-mail address: dipali.koshti@spsu.ac.in

https://doi.org/10.37934/araset.42.1.144157

understanding of features to answer the above question correctly. For better results, it is necessary to co-attend both image and text and truly understand the visual content. Most of the visual question-answering models follow a three-step process: feature extraction, Joint embedding, and answer generation. Different approaches have been experimented with for achieving each of the steps in the literature. For image feature extraction, recent deep learning Models like VGGNet [1-6], ResNet [7-9], and even F-RCNN [10-12] have proved to be a boon. For question feature extraction, literature explored various word embedding techniques ranging from simple word2vec [13,14], and Glove [15,16] to complex LSTM, GRU [14,15,17], and transformers [18-20]. Due to the promising results in language processing, all recent work on VQA makes use of the transformer for both image and question feature extraction [19]. Although using transformers makes the model heavy and complex, with the availability of high computing processing (GPUs) it’s possible to make fully transformer-based VQA models [18-20]. Many authors tried to integrate external knowledge into the VQA framework to enable the model to answer open-domain questions. Open-domain questions require additional outside information besides what is present in the image and question. A number of knowledge bases like WebChild, Wikidata, Wordnet, and DBPedia are freely available at our disposal. Now, the model has to fuse three modalities – image features, question features, and knowledge features. This multi-modality further poses the challenge of integrating and performing multi-modal reasoning. It is clear that integrating external knowledge takes the VQA model one step further and allows it to answer open-domain questions. Another challenge in this domain is the availability of domain-specific datasets. Most commonly used datasets like COCO QA, VQA, and DAQUAR contain images for day-to-day life. So, the models trained on these datasets do not fit well for specific domains such as medical, satellite images or sports images etc. Creating a domain-specific dataset is still a challenge. Few attempts have been made to work on VQA for data visualization [22], arts and paintings [23], satellite images [24], and medical images [25,26]. We attempt to develop a VQA model for school children, especially for educational purposes. The EDUVQA model is not only trained and tested on the dataset created for education but also integrates outside knowledge in terms of facts. Our main contribution is as follows.

i. We propose a new benchmark dataset for image–question answering specially created for educational purposes.
ii. We propose a novel SVO detector that is capable of mining image-question-related facts from the vast fact base.
iii. Finally, we propose the entire multi-modal VQA framework that integrates these facts with the VQA model to generate an answer.

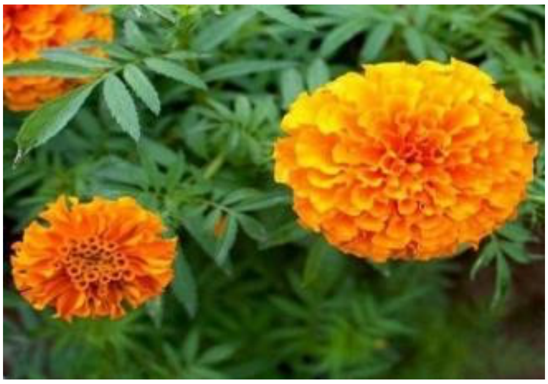


Fig. 1. Question: What is the name of this flower? Answer: marigold

2. Related Work

Many approaches to VQA have been explored in the past. For example, [4,5,27,28] used attention-based mechanisms to solve VQA problems. Instead of attending to every detail of the image and every word of the inquiry, an attention mechanism enables us to concentrate solely on the most crucial portions of the image and questions. [5] proposed a question-driven object-centric attention model called FDA. They extracted global image features using CNN and also extracted local object features (only those objects which are associated to the question). The question is encoded using the LSTM network. [27] divided each object into individual picture crops. A different CNN model is then given to each crop. Instead of extricating attributes from the entire image at once that explain both the objects and their spatial information, the CNN can now concentrate on detecting specific items. Each cropped image also contains location and scale information. They ran a query on digits using the MNIST VQA dataset. [29] presented a co-attention mechanism for VQA that combines image-guided question attention with question-guided visual attention. Both parallel co-attention and alternating co-attention were proposed as attentional mechanisms. By assessing the similarity between the image and question features at all pairs of image places and question locations, they connect the image and question while paying concurrent co-attention. Much of the recent literature [18-20] explores the power of transformers to model a language and vision model. Sometimes, to solve a VQA task, not only the image and question attributes but also the external knowledge is required to answer open-domain questions. Open domain questions are those where the answer to the given question may not be answered just by observing the image and question alone. In such cases, we may need to refer to the outside knowledge bases. For example [15,20,30] uses ConceptNet as an external knowledge base. Other knowledge bases that are available at our disposal are Wikidata [31], DBPedia [32], and Webchild [33]. These external knowledge bases are ready to use and are available for integration with any system.

Most of the above work discussed uses a benchmark dataset like VQA, COCO-QA, and DAQAUR that is composed of day-to-day life images. There is still a challenge to apply VQA to a specific domain, such as Medical, Satellite images, or education due to the lack of enough domain-specific datasets. Many authors created their own domain-specific datasets [34-36]. In [26] authors introduced a dataset for radiology images with correct answers. [26] introduced the first-ever VQA dataset for a medical domain with 4500 training images and question-answer pairs, 500 validation images, and 500 testing images. In [36] authors focused on developing a VQA model for data visualization by proposing a novel FigureQA dataset containing images of bar charts, Pie charts, line graphs, etc. Also, most of the datasets like VQA, COCO-QA, and DAQUAR do not contain any support for external and common-sense knowledge. To address this issue authors of [37] introduced a novel dataset called FVQA that contains questions requiring deeper understanding and reasoning. The dataset includes facts along with the image, question, and answer. Few attempts have been made to improve the learning experience of learners through technology [38-40]. In [39] authors developed a graphical user interface to solve non-linear equations and find roots using numerical methods such as Bisection, Newton's, and secant methods to generate the approximate roots. [40] developed a Figee card using Augmented Reality technology that helps students identify functional groups in the field of organic chemistry.

We address the two challenges of VQA here: The first one is creating a domain-specific dataset. We have chosen education as our domain. Second, we provide a fact base to create an explainable VQA model. This basically helps in answering open-ended questions requiring external knowledge. We propose a complete Multimodal VQA framework with an attention mechanism along with an integrated SVO detector to improve the accuracy and performance of previous works as they focus on the words that are important to match relevant information in our fact relations.

3. Methodology

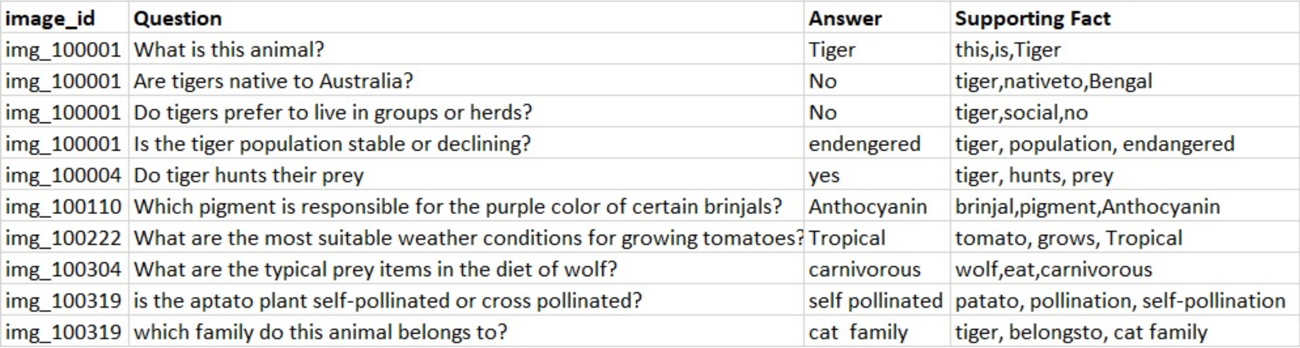
We propose a novel knowledge-incorporated multi-modal VQA model for educational purposes called EDUVQA. The proposed framework is shown in Figure 2. The proposed model is Resnet50 and BERT-based model which integrates the SVO detector to improve the accuracy of the existing fusion mechanism and understand the relationship between sequential elements even if they are far out.

In our baseline model, we used the Resnet50 and RoBERT encoder without using the SVO detector. One of the main drawbacks of VQA models without an SVO detector is that they may generate incorrect or unreliable answers that are not factually correct. This is because these models rely solely on the training data and the learned associations between the visual and textual features to generate answers, without explicitly verifying their factual correctness. In the second model, we integrate the SVO detector into an EDUVQA framework which resulted in an increase of 10% accuracy as compared to the baseline model.

The field of natural language processing has actually seen promising outcomes with BERT-based transformers. By giving it pairs of questions and related images and training it to predict the right answer, the BERT-based transformer can be improved on a VQA job. During inference, the model can be used to generate answers to new questions by encoding the question and the image using the transformer and predicting the answer based on the generated representations. We first discuss the method adopted for the creation of the educational dataset and then the working of the SVO detector model before discussing the proposed VQA framework.

3.1 Dataset Creation

We selected an educational domain to work on VQA. Since the domain is very vast, we restrict our dataset to the science domain with four categories of images: animals, plants, vegetables, and fruits. We propose a new dataset containing 1500 educational images from above mentioned four categories. There are 12000 questions, 750 total unique facts, and 1200 unique answers. We manually created 08 questions for each image and corresponding answers with their IDs for each question. These questions and answers are converted to JSON format. In each category, we have 15 different images for each object (e.g., 15 different images of Tomato, 15 different images of Brinjal, etc.). Each image has been assigned a unique Image ID in the format 400001. To each image, we have assigned the question ID in the form ‘400001001’ where prefix 400001 represents the image id and suffix 001 represents a question number. So, we can frame 625 questions for each image. Answer ID has been assigned in a format of ‘4000010011’ where the first 7 digits ‘400001’ represent the image id, and next 3 digits ‘001’ represents the question number, and the last digit ‘1’ represents the answer id. So, the answer id for the next question can be represented as 4000010021, 4000010031, etc. Also, we created a fact base (knowledge base) in the form of relations with their corresponding fact ids. For the given image, question, and answer, the fact ID in our data set is represented as f_4000010011. For example, one possible fact for the question “What is the habitat of Tiger?” could be < Tigers, belongs To, wild>. The fact is in the form of (subject, verb, object) triplet. Figure 2 shows the sample of our EDUVQA dataset.



image_id	Question	Answer	Supporting Fact
img_100001	What is this animal?	Tiger	this,is,Tiger
img_100001	Are tigers native to Australia?	No	tiger,nativeto,Bengal
img_100001	Do tigers prefer to live in groups or herds?	No	tiger,social,no
img_100001	Is the tiger population stable or declining?	endengered	tiger, population, endangered
img_100004	Do tiger hunts their prey	yes	tiger, hunts, prey
img_100110	Which pigment is responsible for the purple color of certain brinjals?	Anthocyanin	brinjal,pigment,Anthocyanin
img_100222	What are the most suitable weather conditions for growing tomatoes?	Tropical	tomato, grows, Tropical
img_100304	What are the typical prey items in the diet of wolf?	carnivorous	wolf,eat,carnivorous
img_100319	is the aptato plant self-pollinated or cross pollinated?	self pollinated	patato, pollination, self-pollination
img_100319	which family do this animal belongs to?	cat family	tiger, belongsto, cat family

Fig. 2. Sample of EDU-VQA dataset

Table 1 lists the statistics of the EDUVQA dataset. The Longest question in our dataset contains 25 words and the longest answer is four words. The average question length is 13.25. There are 1200 unique answers and 750 unique facts in a database.

Table 1
Analysis of the dataset

# of Images	# of question-answer pair	# of unique answers	# of unique Facts
1500	12000	1200	750

3.2 The SVO Detector

Retrieving a correct supporting fact is essential for generating a correct answer. The objective of an SVO detector is to extract the top 'k' relevant facts in the form of triplet SVO from the fact base given a pair of an image and a question. We train an SVO detector model that accepts a question and an image as inputs and outputs the corresponding {subject, verb, object} triplet. Image features have been extracted using ResNET50. RoBERTa has been used to extract question features. By integrating the picture and question embeddings in the shared space, where bilinear pooling is used for the fusion technique of two modalities, a joint semantic feature embedding is learned. These fused features are then given to three classifiers: subject, relation, and object. Figure 3 depicts the SVO model architecture.

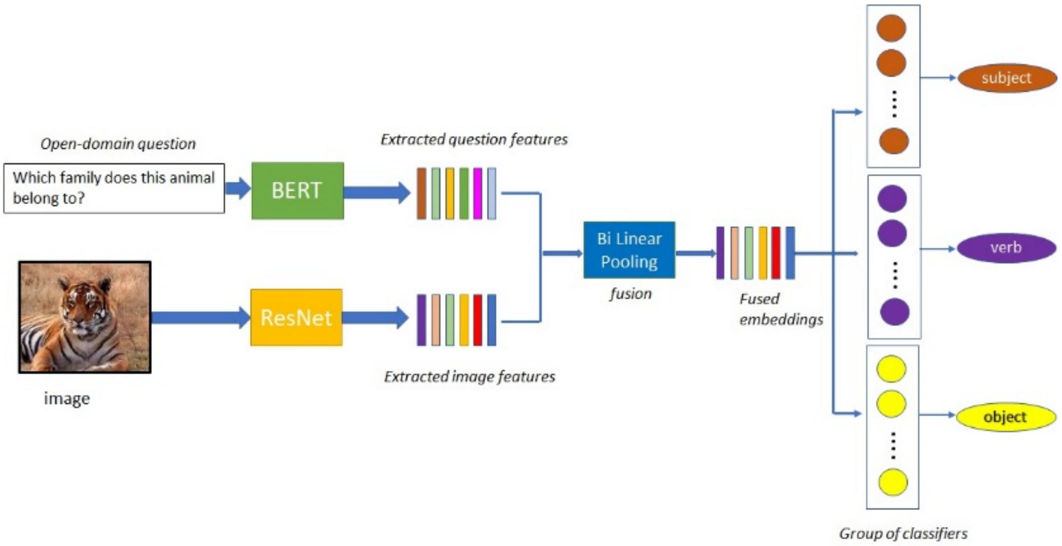


Fig. 3. SVO Detector Model

We used RESNET50 to extract the image representation Given an image ‘I’. The features are extracted from the last convolution layer of the RESNET50 followed by added mean pooling layer to extract dense image representation.

$$v = Meanpool\ (ResNet50(I)) \tag{1}$$

Similarly, to encode a question we use a pre-trained BERT transformer.

$$q\ =\ bert(Q) \tag{2}$$

We apply the linear transformation on Vectors v and q and then apply a non-linear function tanh, respectively. This process encodes the picture and query in a common space.

$$f_v\ =\ tanh\ (w_v v\ +\ b_v\ ) \tag{3}$$

$$f_q\ =\ tanh\ (w_q q\ +\ b_q\ ) \tag{4}$$

Where $W_v, b_v,\ W_q\ and\ b_q$ are learnable parameters for linear transformation. By combining the question and visual embeddings in the shared space via bilinear pooling, a fused semantic feature embedding is learned [38].

$$H\ =\ MLB(f_v\ +\ f_q) \tag{5}$$

The group of linear classifiers is then given these fused representations so as to predict the relation, object and subject in a related fact.

$$Prob(Sub)\ =\ Softmax\ (W_{hs}h\ +\ b_s\ ) \tag{6}$$

$$Prob(verb)\ =\ Softmax\ (W_{hv}h\ +\ b_v\ ) \tag{7}$$

$$Prob(obj)\ =\ Softmax\ (W_{ho}h\ +\ b_o\ ) \tag{8}$$

Where Prob(i) denote the classification probability of parameter i; i = {subject, verb, object}
The Loss function is the summation of loss of all three classifiers.

$$L_{SVO}\ =\ \delta_s\, Loss(s, \hat{s}\ )\ +\ \delta_v\, Loss(v, \hat{v}\ )\ +\ \delta_o\, Loss(o, \hat{o}\ ) \tag{9}$$

Where s,v and o are the actual subject, verb, and object and $\hat{s},\ \hat{v}\ ,\ and\ \hat{o}\ a$re predicted ones respectively. We obtain the optimized value of δ using grid search on dataset. These hyperparameters were discovered using grid search on the development set: δ_s = 1.0, δ_v = 0.8, and δ_o = 1.2. For multi-class classification, the cross-entropy criterion function is indicated by the letter L.

3.3 The Proposed EDU-VQA Architecture

The suggested system intends to create a comprehensive visual framework for question-answering that can reason about open-ended inquiries linked to educational questions and produce responses in a natural language. The EDU-VQA model learns visual and semantic knowledge from the inputs of a semantic question and an image to anticipate the right response. The following are the key elements of the framework we propose shown in Figure 4:

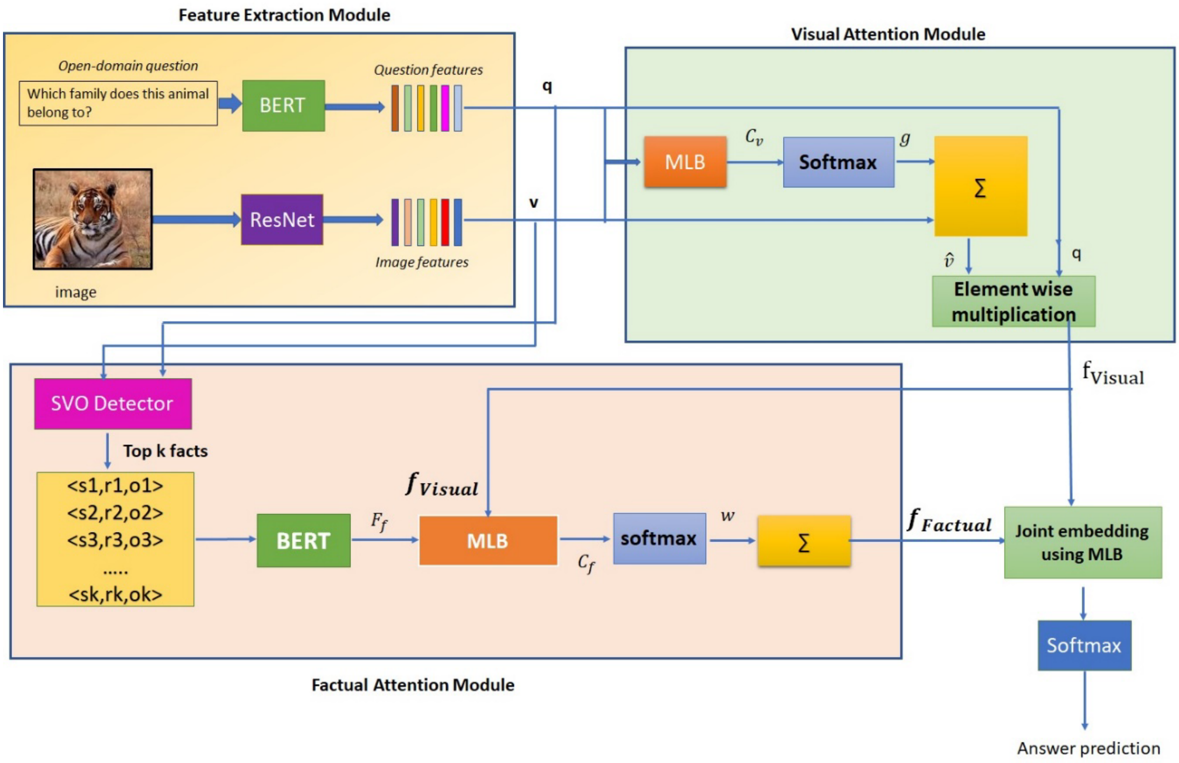


Fig. 4. (a) EDU-VQA Framework Feature extraction module, (b) Visual attention module, (c) Factual attention module, (d) Joint Embedding module, (e) Answer generation module

3.3.1 Feature extraction module

This module is responsible for image and question encoding.

i. **Image encoding**: For image features extraction the pre-trained neural network RESNET is used. The image has been downsized to 224x224 for a more accurate depiction, and the image's feature dimensions are 2048x14x14. Each feature vector f_I corresponds to a 14x14 pixel region of the input images. Features of each region are represented by size 2048. These features may then also be projected on a fully connected layer in order to transform the dimension size to better align with the question features, as done in the case of the ResNet features.

ii. **Question encoding**: Extraction of meaningful text features is done using BERT. Text extraction is done using a pre-trained transformer RoBERTa. We first tokenize the question using the RoBERTa tokenizer, which breaks the question into sub-word units that can be processed by the model. The tokenized question is then given to RoBERTa's layers to output a contextualized representation of the question. This contextualized representation is then combined with the image features, to generate an answer.

3.3.2 Visual attention module

The visual attention module's purpose is to pick out parts of images that are related to the input query and to collect visual semantic representations of those parts. Once image features ‘v’ and question features ‘q’ have been extracted, these two features are fused using Multi modal Bilinear pooling (MLB) [16].

$$C_v \ = \ MLB(v, q) \tag{10}$$

Where the question and image context material are both present in the context vector, ‘C_v’. This Context vector C_v is then given to a linear transformation layer and then a Softmax layer is applied to convert it into attention weights g.

$$g = Softmax\left(W_{c_v} C_v + b_{C_v}\right) \tag{11}$$

where g is of size 14x14.

The weighted sum of all picture regions is used to determine the attended visual characteristics.

$$\hat{v} \ = \ \sum_{i=1}^{14x14} g(i) \ v(i) \tag{12}$$

The attended question and visual vectors are then fused together with a simple element-wise multiplication to get the final visual representation $\mathrm{f_{Visual}}$

$$\mathrm{f_{Visual}} \ = \ \hat{\mathrm{v}} \ \mathrm{X} \tanh(\mathrm{w_q q} \ + \mathrm{b_q}) \tag{13}$$

3.3.3 Factual attention module

The Factual attention module is responsible for selecting important facts with respect to the attended visual features. It basically weighs detected relevant facts by the attended visual features.

Fact detection and encoding: As discussed in section 3.2, given image features(v) and question features (q) our SVO detector outputs the most relevant ‘k’ SVO triplets from the fact base. Let these facts be represented as $\{L_1 , L_2, L_3, \ldots, L_k\}$.

Where $L_i \ = \ \left(s_i , v_{i\,,} r_i\right) , i \ = \ 1,2,3 \ldots\ldots, k$

The next step is to encode these facts. We use BERT encoder to encode these facts similar to the question encoding. The BERT encoder’s output is represented as,

$$F_{fi} \ = \ BERT\ (L_i) \tag{14}$$

And $F_f \ = \ \left[F_{f1} , F_{f2}, \ldots\ldots, F_{fk}\right]$

These facts are then combined with $fvisual$ (attended visual features) by using MLB thus generating factual context vector C_f. Context vector C_f is then converted to attention weights W by passing it through a linear transformation and then applying a Softmax layer.

$$W \ = \ Softmax\left(W_{C_f} \, C_f \ \ + \ b_{C_f}\right) \tag{15}$$

where W is 14x14.

The attended factual features are calculated as weighted sum of representation of all facts. It provides semantic knowledge information for responding to visual queries.

$$f_{Fact} = \sum_{i=1}^{14x14} w(i)\, F_f(i) \tag{16}$$

3.3.4 Joint knowledge embedding

We use bilinear pooling [16] to join the image feature and question feature embeddings. We use early fusion embedding as opposed to late fusion. The early fusion joint embedding combines visual, textual, and factual embeddings before feeding them to a single classifier. In particular, it concatenates the visual and textual embeddings and then applies a fully connected layer to obtain a joint visual-textual embedding. We take the joint visual-textual embedding $f v$isual and factual attended features fFact. To jointly learn visual and semantic knowledge, we combine these two representations using low ranking bilinear pooling [16] and ReLU non-linear activation function.

$$h = MLB(f_{visual} + f_{Factual}) \tag{17}$$

3.3.5 Answer generation

Since VQA is a classification task the top 2 facts matching with the question id were extracted. We define a multi-modal fusion network that combines information from image, text, and fact embeddings. During an evaluation, the trained model is applied to the test dataset. For each input image and question pair, the model outputs a probability distribution over the possible answers. The predicted answer is nothing but the answer with the maximum probability for that input. To evaluate the model performance, the predicted answers are then compared against the actual answers. Similar to other literature we also treat VQA as a multi-class classification problem using the joint embedding learning model, which can jointly encode visual and factual knowledge. A linear classifier is used to deduce the final response and generate the best response.

$$pans = \text{softmax}\ (Wa\text{h} + ba) \tag{18}$$

4. Result and Analysis

4.1 Experimental Setup

The proposed model has been developed using TensorFlow and PyTorch. The model was trained using Google Colab due to its dataset limits and ability to run notebooks in the background. All the models were validated on the validation set after every epoch using the evaluation metric. The evaluation metric used is the overall accuracy, which is the ratio of correctly predicted labels to the total number of labels. The validation set size is set to 20% of the total training data, which is a common practice. The model's performance is evaluated on the validation set after each epoch, and the validation loss is used to determine when to stop training.

4.2 Results

To study the effect of the inclusion of facts in the VQA model, two models were evaluated. In our baseline model, we used the Resnet50 and RoBERT encoder without using the SVO detector. One of the main drawbacks of VQA models without an SVO detector is that they may generate incorrect or unreliable answers that are not factually correct. This is because these models rely solely on the training data and the learned associations between the visual and textual features to generate answers, without explicitly verifying their factual correctness. In the second model, we integrate the SVO detector into a VQA framework which resulted in an increase in accuracy of 10% as compared to the baseline model Figure 5. Shows some of the results obtained for our EDUVQA model.

4.3 Evaluation of SVO Detector

We evaluate the correctness of our SVO detector as the final generated answer may depends on the SVO triplets detected. The evaluation metrics recall@1, recall@5 and recall@10 were calculated (Table 2) similar to [38]. The percentage of numbers for which the correct fact is anticipated in the top k projected facts is known as recall@k. The AdamW optimizer is used to train the detector, with an initial learning rate of 5 × 10^{-5}, and a weight decay of 1 x 10^{-5}. The batch size used was 32, and dropout layer is added before every linear transformation layer. Table 2 shows that both vision and question are important for detecting SVO triplet. Also, question is more important than vision as question only model predicts subject, verb and object more accurately.

Table 2
Results of SVO detector

	Accuracy			**Fact (Recall)**		
	Subject	Verb	Object	Recall@1	Recall@5	Recall@10
Question only	83.4	71.34	78.02	65.76	81.24	79.04
Visual only	74.56	65.35	81.03	63.4	78.9	75.03
Vision+Question	98.2	78.3	84.8	69.9	85.1	87.6

4.4 Evaluation of EDUVQA Model

For question encoding question, the vector size of 768 is used. For facts encoding the top two facts were generated. All image representations are in the form of vectors of size 2048. We implement our model using PyTorch framework. In our experiments, we utilize the AdamW optimization for the training process with batch size of 64, an initial learning rate of 5 × 10^{-5}, and a weight-decay of 1x10^{-5}. If the validation accuracy does not improve then early stopping is applied at the last three validations. Table 3 shows the accuracy of two models, one without an SVO detector and another with SVO detector integrated in the model.

Table 3
Comparison of Proposed VQA Models

Total Epoch	Model output		
	Model	Training Accuracy	Validationaccuracy
40	EDU-VQA Without SVO detector	73.67	70.03
40	EDU-VQA With SVO detector	82.45	81.24

From Table 3, we observe that the performance of the model with a SVO detector significantly increasing accuracy as compared to that of without SVO detector, especially for questions that require knowledge beyond what is provided in the image and text inputs. With a fact detector, the model is able to extract factual information from external sources and use it to improve its predictions.

4.5 Ablation

We undertake four ablation experiments to investigate the function of various model elements. The ablation outcomes of comparing baseline models that are trained on the training set and assessed on the validation set are shown in Table 4. The ablation experiments specifically are as follows.:

Table 4
Ablation of different components of the model

Method	Accuracy
Ques + Img	69.35
Ques + facts	65.45
Ques+Img + Visual attention	70.03
Ques + Img + Visual attention + Fact attention	81.24

Ques +img: This is very basic model where only image and question features are considered.

Ques + facts: This is the model where only question and fact is considered without considering visual content.

Quest + Img + Visual attention: This is the model where all three modalities along with visual attention is applied.

Quest + Img + Visual Attention + Fact attention: The final model where all three modalities along with both visual and factual attention is applied.

Q-Which family is the tiger categorized in?
Detected facts: <tiger, belongsTo, cat family>
<tiger, IsA, largest cat>
Ground truth: cat family
Predicted: cat family

Question: where does Patato grow ?
Detected Fact: <patato, grows, underground>
: <patato, grows, belowground>
Ground truth: under the ground
Predicted: Below ground

What is the enzyme in garlic that may help to prevent blood clots?
Detected Facts: <Garlic, contains, sulfur-based enzyme>
<Garlic, reduce, cholesterol>
Ground Truth: sulphur-based enzyme
Predicted: sulphur-based enzyme

What is the taxonomic name for the object shown in image?
Detected Facts: <carrot, scientific name, Daucus carota>
<carrot, called, Daucus carota>
Ground truth: Daucus carota
Predicted: Daucus carota

Fig. 5. Sample results of EDUVQA model

4.6 Comparison with State-of-the-Art

To evaluate the effectiveness of our dataset, we evaluate KB-VQA [28], MCAN [41] and KRISP [32] models on our proposed dataset. It is clear from the results shown in Table 5 that our dataset is more challenging and requires way of extraction of correct relevant facts from the fact base. The KBVQA model [28] uses ConceptNet as an external knowledge to extract outside information. This model reports 78.01 % accuracy on our dataset, this clearly shows that just extraction of external knowledge from Conceptnet is not enough for such domain specific questions rather more relevant information is needed that is present in the form of facts in our database.

Table 5
Comparison with state-of-the-art models on EDUVQA dataset

Method	Overall Accuracy on proposed dataset
KBVQA [28]	78.01
MCAN [41]	79.04
KRISP [32]	76.7
EDUVQA	81.24

5. Conclusion and Future Work

We address the issue of the domain-specific dataset for education and hence propose a novel dataset containing image, question, and answer along with the fact. Our fact base contains more relevant information needed to answer education-related questions as compared to the information extracted from the publicly available knowledge bases. We propose a novel SVO detector model capable of finding facts related to the given image and question pair. To check the effectiveness of our SVO detector, we developed two models of EDUVQA – one without an SVO detector and one with an integrated SVO detector in the VQA model. Our study reveals that the inclusion of SVO detector allows the model to integrate external knowledge and answer more complex questions. Moreover, attention-based mechanism – visual and factual attention improves the model performance. Future work includes the expansion of dataset for more objects like various plants, creepers etc. Also, the model can be further improved by applying more complex co-attention mechanism as we used a simple bilinear pooling method for joint knowledge embeddings.

Acknowledgement
This research was not funded by any grant.

References

[1] Antol, Stanislaw, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. "Vqa: Visual question answering." In *Proceedings of the IEEE international conference on computer vision*, pp. 2425-2433. 2015. https://doi.org/10.1109/ICCV.2015.279

[2] Ruwa, Nelson, Qirong Mao, Liangjun Wang, and Ming Dong. "Affective visual question answering network." In *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*, pp. 170-173. IEEE, 2018. https://doi.org/10.1109/MIPR.2018.00038

[3] Malinowski, Mateusz, Marcus Rohrbach, and Mario Fritz. "Ask your neurons: A neural-based approach to answering questions about images." In *Proceedings of the IEEE international conference on computer vision*, pp. 1-9. 2015. https://doi.org/10.1109/ICCV.2015.9

[4] Gu, Geonmo, Seong Tae Kim, and Yong Man Ro. "Adaptive attention fusion network for visual question answering." In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 997-1002. IEEE, 2017. https://doi.org/10.1109/ICME.2017.8019540

[5] Ilievski, Ilija, Shuicheng Yan, and Jiashi Feng. "A focused dynamic attention model for visual question answering." *arXiv preprint arXiv:1604.01485* (2016).

[6] Kamoji, Supriya, Mukesh Kalla, and Insiya Shamshi. "A Framework for Flood Extent Mapping using CNN Transfer Learning." *International Journal of Intelligent Systems and Applications in Engineering* 10, no. 3s (2022): 150-157.

[7] Kafle, Kushal, and Christopher Kanan. "Answer-type prediction for visual question answering." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4976-4984. 2016. https://doi.org/10.1109/CVPR.2016.538

[8] Nguyen, Duy-Kien, and Takayuki Okatani. "Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6087-6096. 2018. https://doi.org/10.1109/CVPR.2018.00637

[9] Gupta, Deepak, Pabitra Lenka, Asif Ekbal, and Pushpak Bhattacharyya. "A unified framework for multilingual and code-mixed visual question answering." In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pp. 900-913. 2020.

[10] Yang, Chao, Mengqi Jiang, Bin Jiang, Weixin Zhou, and Keqin Li. "Co-attention network with question type for visual question answering." *IEEE Access* 7 (2019): 40771-40781. https://doi.org/10.1109/ACCESS.2019.2908035

[11] Gao, Lianli, Liangfu Cao, Xing Xu, Jie Shao, and Jingkuan Song. "Question-Led object attention for visual question answering." *Neurocomputing* 391 (2020): 227-233. https://doi.org/10.1016/j.neucom.2018.11.102

[12] Lu, Jiasen, Dhruv Batra, Devi Parikh, and Stefan Lee. "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks." *Advances in neural information processing systems* 32 (2019).

[13] Shih, Kevin J., Saurabh Singh, and Derek Hoiem. "Where to look: Focus regions for visual question answering." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4613-4621. 2016. https://doi.org/10.1109/CVPR.2016.499

[14] Zhang, Peng, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. "Yin and yang: Balancing and answering binary visual questions." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5014-5022. 2016. https://doi.org/10.1109/CVPR.2016.542

[15] Zhang, Liyang, Shuaicheng Liu, Donghao Liu, Pengpeng Zeng, Xiangpeng Li, Jingkuan Song, and Lianli Gao. "Rich visual knowledge-based augmentation network for visual question answering." *IEEE Transactions on Neural Networks and Learning Systems* 32, no. 10 (2020): 4362-4373. https://doi.org/10.1109/TNNLS.2020.3017530

[16] Kim, Jin-Hwa, Jaehyun Jun, and Byoung-Tak Zhang. "Bilinear attention networks." *Advances in neural information processing systems* 31 (2018).

[17] Yang, Chao, Mengqi Jiang, Bin Jiang, Weixin Zhou, and Keqin Li. "Co-attention network with question type for visual question answering." *IEEE Access* 7 (2019): 40771-40781. https://doi.org/10.1109/ACCESS.2019.2908035

[18] Su, Weijie, 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).

[19] Tan, Hao, and Mohit Bansal. "Lxmert: Learning cross-modality encoder representations from transformers." *arXiv preprint arXiv:1908.07490* (2019). https://doi.org/10.18653/v1/D19-1514

[20] Gardères, François, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lecue. "Conceptbert: Concept-aware representation for visual question answering." In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 489-498. 2020. https://doi.org/10.18653/v1/2020.findings-emnlp.44

[21] Li, Xiujun, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang *et al.,* "Oscar: Object-semantics aligned pre-training for vision-language tasks." In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pp. 121-137. Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-58577-8_8

[22] Kafle, Kushal, Robik Shrestha, Scott Cohen, Brian Price, and Christopher Kanan. "Answering questions about data visualizations using efficient bimodal fusion." In *Proceedings of the IEEE/CVF Winter conference on applications of computer vision*, pp. 1498-1507. 2020. https://doi.org/10.1109/WACV45572.2020.9093494

[23] Garcia, Noa, Chentao Ye, Zihua Liu, Qingtao Hu, Mayu Otani, Chenhui Chu, Yuta Nakashima, and Teruko Mitamura. "A dataset and baselines for visual question answering on art." In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pp. 92-108. Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-66096-3_8

[24] Lobry, Sylvain, Diego Marcos, Jesse Murray, and Devis Tuia. "RSVQA: Visual question answering for remote sensing data." *IEEE Transactions on Geoscience and Remote Sensing* 58, no. 12 (2020): 8555-8566. https://doi.org/10.1109/TGRS.2020.2988782

[25] Liu, Shengyan, Xuejie Zhang, Xiaobing Zhou, and Jian Yang. "BPI-MVQA: a bi-branch model for medical visual question answering." *BMC Medical Imaging* 22, no. 1 (2022): 1-19. https://doi.org/10.1186/s12880-022-00800-x

[26] Ben Abacha, Asma, Mourad Sarrouti, Dina Demner-Fushman, Sadid A. Hasan, and Henning Müller. "Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain." In *Proceedings of the CLEF 2021 Conference and Labs of the Evaluation Forum-working notes*. 21-24 September 2021, 2021.

[27] Burt, Ryan, Mihael Cudic, and Jose C. Principe. "Fusing attention with visual question answering." In *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 949-953. IEEE, 2017. https://doi.org/10.1109/IJCNN.2017.7965954

[28] Koshti, Dipali, Ashutosh Gupta, and Mukesh Kalla. "Knowledge Blended Open Domain Visual Question Answering using Transformer." In *2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, pp. 823-828. IEEE, 2023. https://doi.org/10.1109/ICAIS56108.2023.10073911

[29] Lu, Jiasen, Jianwei Yang, Dhruv Batra, and Devi Parikh. "Hierarchical question-image co-attention for visual question answering." *Advances in neural information processing systems* 29 (2016).

[30] Li, Guohao, Hang Su, and Wenwu Zhu. "Incorporating external knowledge to answer open-domain visual questions with dynamic memory networks." *arXiv preprint arXiv:1712.00733* (2017).

[31] Shah, Sanket, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. "Kvqa: Knowledge-aware visual question answering." In *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, pp. 8876-8884. 2019. https://doi.org/10.1609/aaai.v33i01.33018876

[32] Marino, Kenneth, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. "Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14111-14121. 2021. https://doi.org/10.1109/CVPR46437.2021.01389

[33] Cao, Qingxing, Bailin Li, Xiaodan Liang, and Liang Lin. "Explainable high-order visual question reasoning: A new benchmark and knowledge-routed network." *arXiv preprint arXiv:1909.10128* (2019).

[34] Lau, Jason J., Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. "A dataset of clinically generated visual questions and answers about radiology images." *Scientific data* 5, no. 1 (2018): 1-10. https://doi.org/10.1038/sdata.2018.251

[35] Ben Abacha, Asma, Mourad Sarrouti, Dina Demner-Fushman, Sadid A. Hasan, and Henning Müller. "Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain." In *Proceedings of the CLEF 2021 Conference and Labs of the Evaluation Forum-working notes*. 21-24 September 2021, 2021.

[36] Kahou, Samira Ebrahimi, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. "Figureqa: An annotated figure dataset for visual reasoning." *arXiv preprint arXiv:1710.07300* (2017).

[37] Wang, Peng, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. "Fvqa: Fact-based visual question answering." *IEEE transactions on pattern analysis and machine intelligence* 40, no. 10 (2017): 2413-2427. https://doi.org/10.1109/TPAMI.2017.2754246

[38] Arifin, Raihan Bt. "Rabbani Education: Facing Realities and Readiness for the Challenges of Future Education." *International Journal of Advanced Research in Future Ready Learning and Education* 30, no. 1 (2023).

[39] Ab Wahab, Nurul Ain, and Mohd Agos Salim Nasir. "Graphical User Interface for Solving Non-Linear Equations for Undergraduate Students." *International Journal of Advanced Research in Future Ready Learning and Education* 30, no. 1 (2023): 25-34.

[40] Suhaimi, Elmi Sharlina Md, Zuhaizi Abdullah, Norazreen Muhamad, Nik Khadijah Nik Salleh, and Ahmad Affendy Abdullah. "FIGEE CARD: Pembelajaran Interaktif Kumpulan Berfungsi Kimia Organik: FIGEE CARD: Interactive Learning of Organic Chemistry Functional Groups." *International Journal of Advanced Research in Future Ready Learning and Education* 30, no. 1 (2023): 13-24.

[41] Yu, Zhou, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. "Deep modular co-attention networks for visual question answering." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6281-6290. 2019. https://doi.org/10.1109/CVPR.2019.00644