EDUVI: An Educational-Based Visual Question Answering and Image Captioning System for Enhancing the Knowledge of Primary Level Students

Manisha Gupta[1], Priya Asthana[2], Preetvanti Singh[3]

[1,2,3]Dayalbagh Educational Institute, Agra, UP, India

ABSTRACT

Within the last several years, the revolution in online education has fundamentally transformed the idea of traditional education.The trend of online education is exploding in popularity in the modern educational system. However, it is challenging for primary-level students to adapt to learning in this novel environment.When enrolled in online courses, students at this level encounter a variety of difficulties.Hence, to provide better learning methods, this research study focuses on developing an education based EDUVI system for primary level students which helps the students in visual learning. The proposed system will help the students for self-learning without any assistance using a simplified and interactive platform. The developed system integrates visual question answering and image captioning system where students can import the image and extract answer based on the query asked by them or can generate the description or caption.

Keywords:Visual learning, Primary School Students, Visual Question Answering, Image Captioning, Computer Vision.

INTRODUCTION

With the advancement in technology, the trend of online learning among students is increasing exponentially. Moreover, every individual has access to a quality education whenever and wherever required. The revolution of online education has altered the concept of traditional education drastically within the last couple of years. According to the most recent survey by Brainly (IANS, 2021) almost 54% of Indian students are comfortable with online education. However, primary-level students found it hard to get used to studying in this new environment. Students of this level face various challenges when they are subjected to online classes. One of the main challenges was to keep the students focused as children were not able to associate schoolwork with their laptops and computers. They use these devices to play games or watch videos and therefore are unable to focus on schoolwork when it is on their computers. Without focus use of online education portals for watching the video lectures or solving the assignments can be much harder.

One way to overcome this is engaging students with the content to make them learn faster, and this is where visual content excels. Visuals draw the attention of the learner and enhance engagement. Visual information plays an important role in education because it is more appealing to students rather than hearing or reading plain text. A visual approach to learning includes use of computer vision to make the learning process more effective because it is believed that learning through visuals is more memorable for longer periods of time as compared to textual learning. Visual learning is a learning style where students prefer to use visuals like images and graphics for communicating thoughts and ideas. Visual learning strategies can help students develop skills such as critical thinking, better decision-making, problem-solving, and better understanding. This can be very useful at the primary school level, particularly in an online learning environment, because it makes communication simpler and quicker, drives motivation, and stimulates emotions.

The Visual Question Answering (VQA) system (Kafle et al., 2017; Zhou et al., 2020), a visual approach to learning, is a powerful system for elaborating the concepts of learning to the students by not only enabling them in understanding the picture and objects but also giving them the knowledge based on the question asked. It is a system that integrates computer vision and natural language processing for stimulating research (Fig.1). Computer vision is used for acquiring, processing, and understanding images with the motive being to teach machines *how to see*. On the other hand, natural language processing is concerned with enabling interactions between computers and humans in natural language, *i.e.* teaching machines *how to read*. Image Captioning (IC), an enhanced version of VQA, is the process of generating a textual description for given images. Image captioning can be viewed as an end-to-end sequence-to-sequence problem where images are converted from a sequence of pixels to a sequence of words.

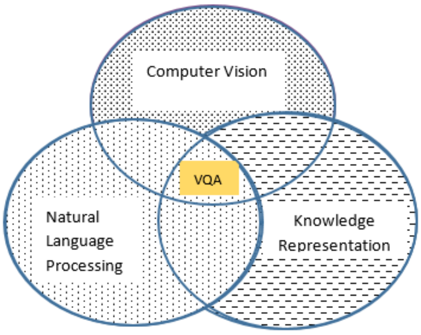


Fig 1: The VQA approach for knowledge representation

The main goal of this paper is to develop a visual learning system for primary school students to improve their learning skills by integrating Visual Question Answering and Image Captioning approaches for high-level scene interpretation. The aim of developing such a system is 2 folds:

a. Increasing their learning ability
b. Enhancing their knowledge in understanding the complex concepts.

Significance of the study

Online education is proved to be very beneficial over traditional education because it provides remote learning, reduced costs, increased course variety and personalized education that further increases the technical skills**.** Despite of various benefits of online education system there are several problems that are faced by students during online learning. Even while the younger generation is adept at using computers, this does not equate to digital literacy. It is quite difficult for some students to learn how to use many pieces of software effectively when using an online learning system as most of them are not familiar with the technology and the system. To make the scope of online education wider, the need is to design a system in a way that can be easily understandable by any individual without any difficulty. This research study is an effort in this direction.

RELATED WORK

VQA is utilized in various domains for educating users. Medical-VQA aims to accurately answer a clinical question presented with a medical image. Bansal, Gadgil, Shah&Verma (2019) created a VQA System that uses medical images as context to generate answers. The VQA pipeline classified the questions into two groups, the first group of questions generated answers from a fixed pool of predefined answer categories and the second group of questions produced answers based on the abnormalities seen in the image. He et al. (2020) created a VQA dataset from pathology images together with a question and their correct answers. A semi-automated pipeline was developed to extract pathology images and captions from textbooks and generate question-answer pairs from captions using natural language processing. Zhan, Liu, Fan, Chen& Wu (2020) developed a conditional reasoning framework for Med-VQA to automatically learn reasoning skills for various Med-VQA tasks. A question-conditioned reasoning module was created to guide the importance selection over multimodal fusion features. Al-Sadi, Al-Ayyoub, Jararweh&Costen (2021) investigated several deep learning approaches in building a medical VQA system based on Image CLEF's VQA-Med dataset. A hierarchical model consisting of many sub-models was created to handle these questions. Chebbi (2021) developed a method for producing visual questions on radiology images using augmentation techniques for extracting features from a picture.

The system was implemented using Tensorflow. Gupta, Suman&Ekbal (2021) developed a hierarchical deep multi-modal network toanalyze and classify end-user queries and incorporated a query-specific approach for answer prediction. A question segregation technique for Med-VQA was created by integrating it with the hierarchical deep multi-modal neural network to generate proper answers to the medical image queries. Liu et al. (2021) presented a large bilingual dataset, SLAKE, with comprehensive semantic labels annotated by experienced physicians and a structural medical knowledge base for Med-VQA. Li et al. (2022) developed a bi-level representation learning model with two reasoning modules for the medical VQA task. A sentence-level reasoning module was created to learn sentence-level semantic representations from multimodal input, and the token-level reasoning module was employed using an attention mechanism to generate a multimodal contextual vector. Wu et al. (2022) introduced mainstream datasets (VQA-RAD, VQA-Med, Path-VQA and SLAKE) used for Med-VQA tasks and elaborated the prevalent methods for Medical VQA tasks.

Surveillance of individuals using visual data requires human-level capabilities for understanding the characteristics that differentiate one person from another. Toor, Wechsler& Nappi (2019) developed a system for biometric-based surveillance by utilizing models that are relevance-aware to triage images and videos based on interaction with single or multiple users. The system focused on the detection of people via their appearance and clothing.Vo, Phung& Ly (2020) created a QA system based on Task Ontology for surveillance to map a question sentence to corresponding tasks for generating the answer. The focus was on Pose estimation/tracking and Skeleton-based action recognition.Wang et al. (2021) developed a method for modeling human–machine collaboration based on digital twins. VQA technology was introduced in this system for consistent integration.

VQA can play an important role in educating tourists. Guerrieri, Ghiani&Manni (2017) developed a Tourist Advisor for providing accurate answers to tourists' questions related to Italian tourist destinations. The System presented a controlled natural language to represent the domain knowledge and common sense. Nugraha&Chahyati (2020) developed a training model for object detection in VQA. YOLO and RetinaNet were used for object detection. Siregar&Chahyati (2020) compiled a Monas VQA dataset that uses Bahasa Indonesia in the question and Monas, a memorial monument for Indonesian, as the image-specific context. VQA was solved using convolutional neural network for image embedding and techniques from natural language processing for sentence embedding. Yang et al. (2020) focused on the needs of QA in the tourism field and constructed a QA system based on the knowledge graph of tourism. An algorithm to identify the tourism entities in question was proposed according to the characteristics of the tourism entities.

The VQA system is also being used in the education domain to support visual learning. He et al. (2017) presented an AI-based robot system of VQA for metacognition tutoring and geometrical thinking training with characteristics of contextual teaching by mining knowledge from the real world directly. For metacognition tutoring objects in the real world were detected and a set of learning materials associated with the objects was presented to learners. For geometrical thinking training, an automatic questioning-and-answering section was employed to engage the learner to think. Ringe, Marathe, Manjrekar & Shetty (2020) presented a web app using VQA for pre-schoolers on their academic journey. Based on the image and a question related to the image as input a natural language answer as an output was generated. Cui, Han & Zhu (2022) presented a VQA-based online teaching effect evaluation model. A guide-attention model was developed to discover the directive clues and the self-attention models were used to reweight the vital feature for locating the critical information on the whiteboard and students’ faces. Lu, Ye, Ren& Yang (2022) introduced a textual distractors generation task for VQA to generate challenging, yet meaningful distractors given the context image, question, and correct answer. The developed task aims at generating distractors without ground-truth training samples.

It can be observed from the above review that VQA systems are gaining popularity for educating the learners. These systems are mainly being used in the medical domain. However, their use in the education domain can also be very useful, particularly for primary (elementary) school students in overcoming challenges during online learning. Thus,an Educational Visual Question Answering and Image Captioningsystem is designed for primary-level students in this study. The motivation behind this research study is to make learning style more interactive and creative for primary-level students by enabling them to accurately answer questions from an image.

THE EDUCATIONAL VISUAL QUESTION ANSWERING AND IMAGE CAPTIONING (EDUVI) SYSTEM

This section presents the developed EDUVI system for primary students to improve their understanding capability. EDUVI system is a multimodal system that integrates both VQA and IC systemsto generate the answer to the question asked for a given image; and to generate the caption for the image. The system takes image and framed question as an input and gives the answer to the query as an output. In case the students are not able to frame the question in accordance with the image because of a complicated image or due to lack of understanding about the image, a caption to the image is generated.

The EDUVI system is first trained for teaching machines to learn how to see and how to read using the COCO and VQA V2 datasets. COCO dataset is used for describing the images as this dataset is trained over 5 captions per image and VQA v2 dataset is used for answering the questions about images which requires knowledge and understanding of visual features. Training was done to perform following CV and NLP tasks (Table 1).

Table 1: Tasks list

Computer Vision (CV) Tasks	
CV Tasks	**Example**
Object recognition	Identify the objects present in the image
Object detection	Are there any fruits in the image?
Attribute classification	What shape is the orange?
Scene classification	Is it sunny/cloudy?
Counting	How many students are in the image?
Activity recognition	Are the young men playing frisbee?
Spatial relationship among the objects	Is there a mushroom in the pizza
Common-sense reasoning	Is this a painting or a real image?
Knowledge-based reasoning	What is the astronaut doing?
Emotion recognition	Is the boy crying?
Stuff image segmentation	What scene does this image show?
Panoptic (Full scene segmentation)	Are the girls going to school on a bicycle?
Natural Language Processing (NLP) Tasks	
NLP Tasks	**Description**
Tokenization	Is the process of splitting a phrase, sentence, or paragraph of text into smaller words.
Part-of-speech tagging	Assigning the parts of speech (verb, noun, adjective) to tokenized words.
Stop word Removal	Is process of removing common words (like articles, prepositions, pronouns, conjunctions).
Named Entity Recognition	Identifies named entities in the text.

EDUVI image dataset was prepared from the images of class 4th NCERT EVS Textbook “Looking Around”. A total of 100 informative images were collected from the book in .jpg format which contains some unique information like animals, birds, aquatic animals, kitchen, park, bridges and so on.The trained dataset was then used to test the EDUVI dataset, i.e. to learn the visual and textual features of the images in EDUVI dataset.

Functionalities of the system

The system supports the following functions:

Answering Question from different categories: The system is capable of answering the questions for each image in following 8 categories (Table 2).

Table 2: Question Answer Categories of the system

S.No.	**Question Categories**	**Description**	**Example**
1.	Verification	States binary answer Yes/No	Are they playing cricket?
2.	Disjunctive	States if certain case is present	Is papaya present in the fruit’s basket?
3.	Concept Completion	Who? What? When? Where?	What is the cat doing?
4.	Feature Specification	Color, Shape, Size like attributes	Which color is the student wearing?
5.	Quantification	How much? How many?	How many persons are there in the image?
6.	Definition	What does an image describe?	Describe the image?
7.	One-word	Returns the answer in one word	What statue is shown in the image?
8.	Sentiment-based	Answers based on emotions	Is the boy crying?

The question categories were chosen on the basis of learning and understanding capability of primary level students because most of the students in their learning age focus on attributes classification, for example, identifying the presence of a specific object in the image or giving description of the image.

Generating Description to the Image: This feature enables the system to provide description of the image if a user is not able to understand the complete image.

Knowledge Development: The system enables the students in developing the knowledge and improving their learning ability as well as thinking skills. For example, if the input image is of any animals and question is asked by the students about what is the type animal class is this, the system provides the answer as "Mammal".

Keyword Based Answer Generation: In order to reduce the complexityof learning, proposed system is capable of generating an answer on the based on the keywords present in the query. If a user is not able to frame the question related to that image then just by specifying the keywords system will generate the answer

Knowledge Correction: The system guides the student by generating correct responses if they face difficulty in understanding the image or make wrong interpretations.

Knowledge generated from anonymous and complex images: The system allows the students to gain knowledge from a complex image which the students may not know.

Knowledge of arbitrary tested image: The EDUVI system can even answer the questions for some random images which are not in the EDUVI dataset.

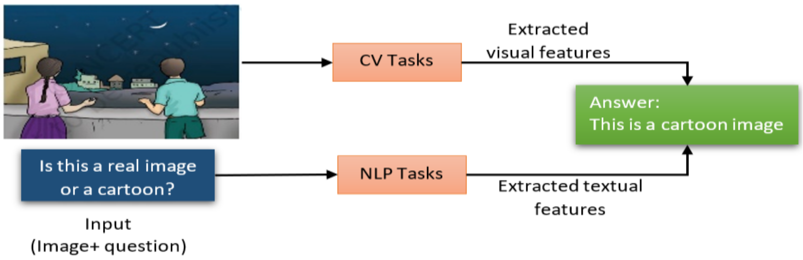


Fig. 3 Workflow of EDUVQA Module

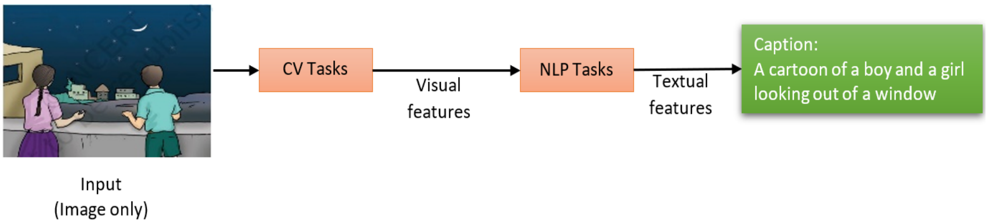


Fig. 4 Workflow of EDUIC Module

Experimentation and Results

The User Interface of the System

The interface for the EDUVI system is shown in Fig. 5. It is an interactive UI, developed in gradio app, where students can easily add an input image and frame a query related to that image. The answer is produced after clicking on submit button.

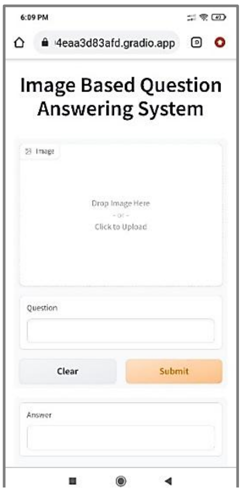


Image Based Question Answering System

Fig. 5 Interface of developed system

For example, the image of fruits and the question "Is papaya present in the basket?" is given as input. After clicking on submit button the answer is generated and displayed as shown in Fig. 6.

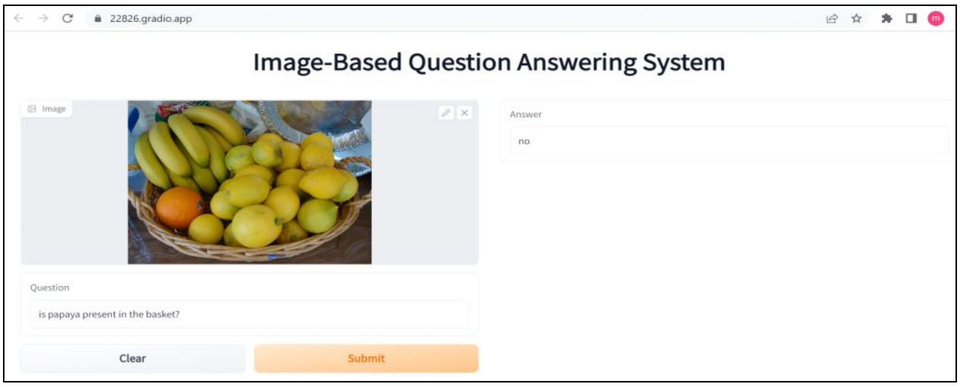


Image-Based Question Answering System

Answer: no

Question: is papaya present in the basket?

Fig. 6 Interface of EDUVQA System

An interface for image captioning system is also shown in Fig. 7 for the input image.The caption is produced after clicking on the submit button.

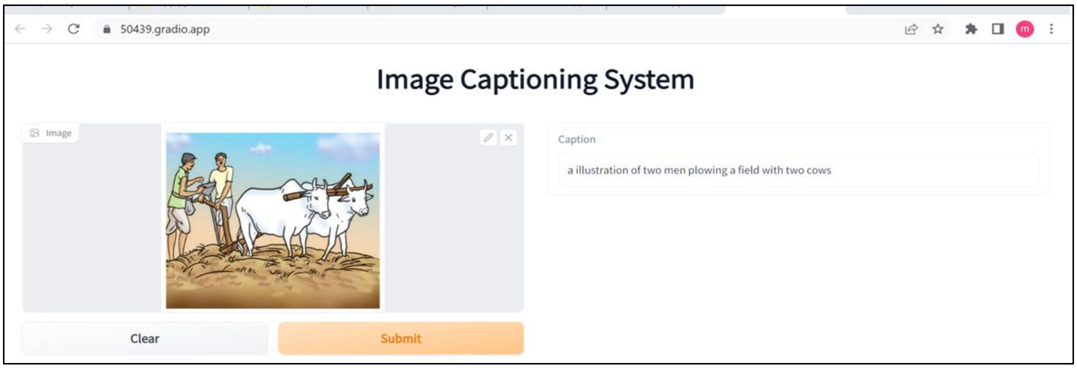


Image Captioning System

Caption: a illustration of two men plowing a field with two cows

Fig. 7 Interface of EDUIC System

Experimentation

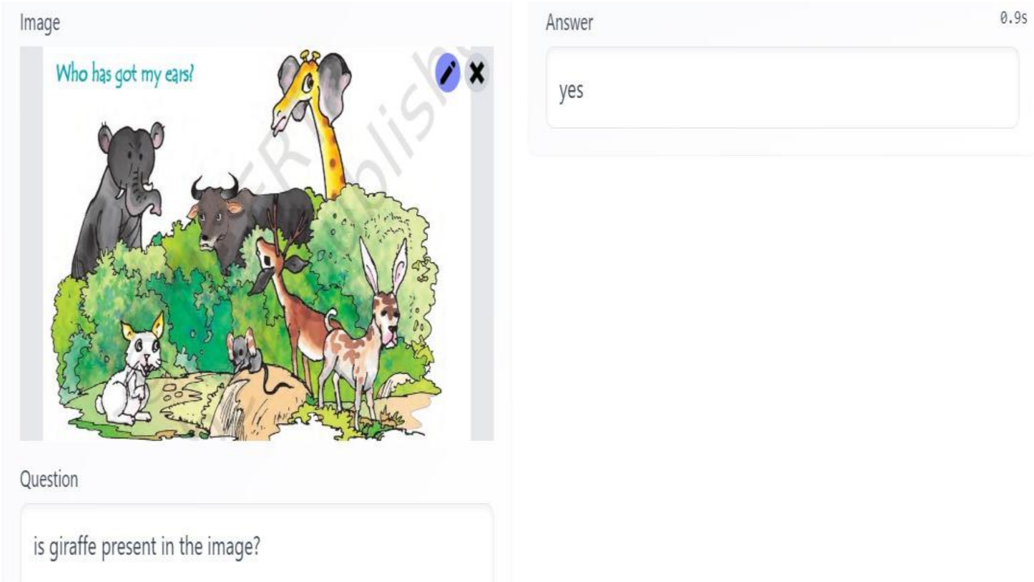
The image dataset of the EDUVI contains 100 images. Few images in .jpg format and the questions (in the form of text)are taken as input to show the applicability of the developed system. The questions from the question categories can be answered using the following CV and NLP tasks given in Table 3:

Table 3: Question categories mapping with CV and NLP tasks

S.No.	Question Category	CV Tasks	NLP Tasks
1.	Verification	Object recognition, Stuff image segmentation, Scene classification	Tokenization, Part-of-speech tagging, Named entity recognition
2.	Disjunctive	Object detection, panoptic	Tokenization, Part-of-speech tagging, Stop word removal, Named entity recognition
3.	Concept Completion	Activity recognition	Tokenization, Part-of-speech tagging, Named entity recognition
4.	Feature Specification	Attribute classification	Tokenization, Part-of-speech tagging, Named entity recognition
5.	Quantification	Counting	Tokenization, Part-of-speech tagging, Named entity recognition
6.	Definition	Spatial relationship, Scene classification	Tokenization, Part-of-speech tagging, Named entity recognition
7.	One-Word	Common sense reasoning	Tokenization, Part-of-speech tagging, Named entity recognition
8.	Sentiment-Based	Emotion recognition	Tokenization, Part-of-speech tagging, Named entity recognition

Answering question in different Categories: It can be seen that question in each question category mentioned in Table 3, is answered by the system.

Verification



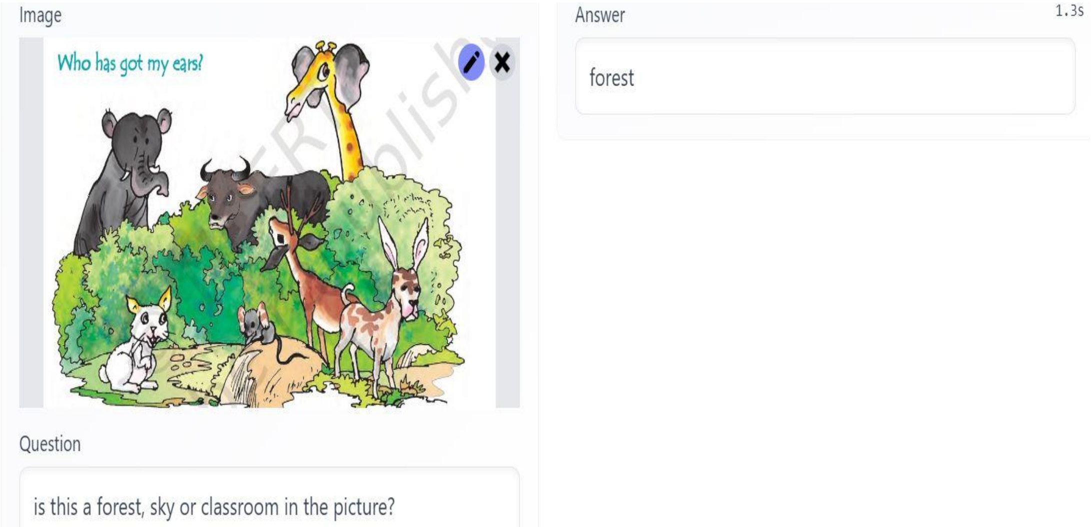
Question

is giraffe present in the image?

Answer

yes

Disjunctive



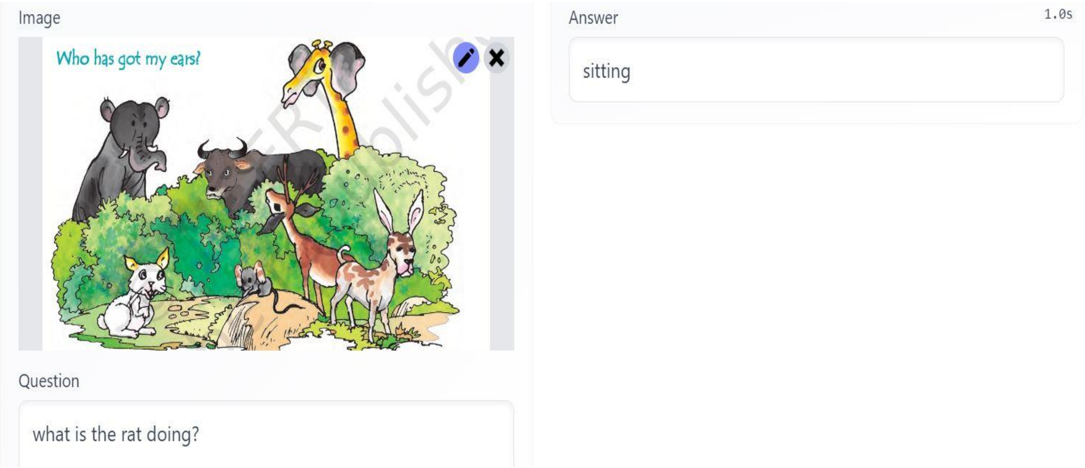
Question

is this a forest, sky or classroom in the picture?

Answer

forest

Concept Completion



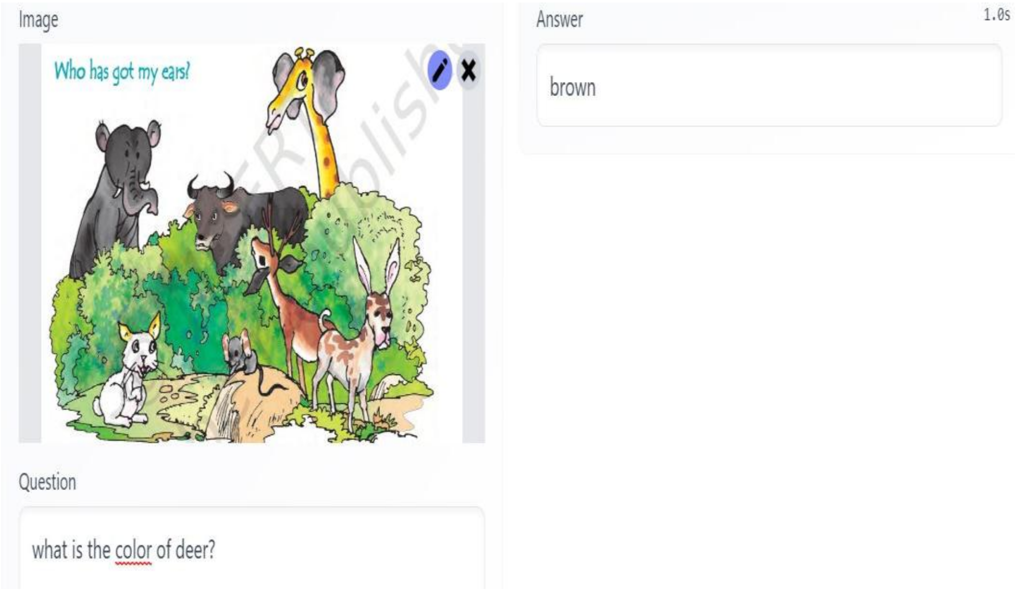
Question

what is the rat doing?

Answer

sitting

Feature Specification



Image

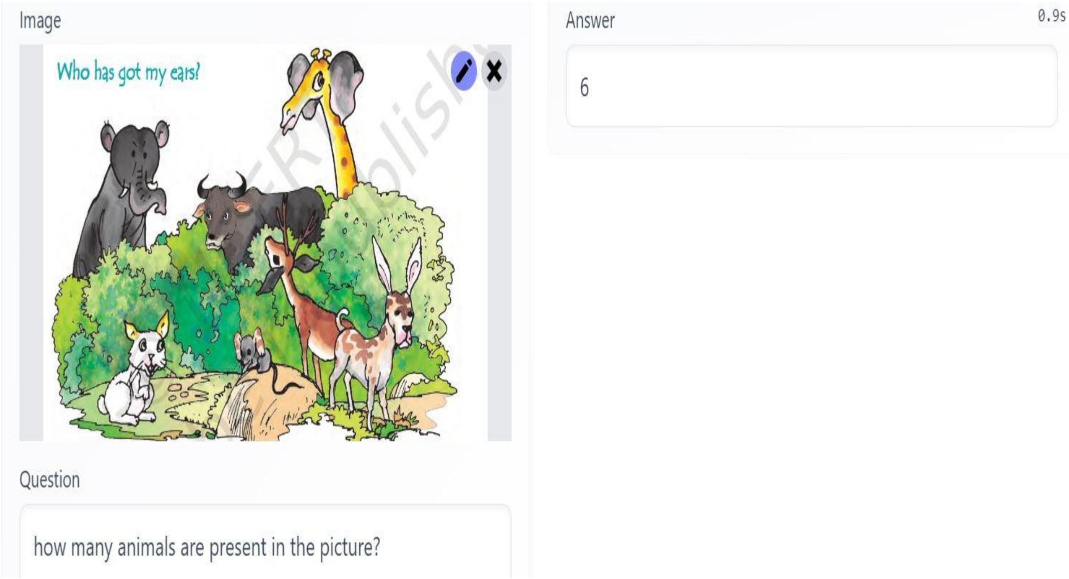
Question

what is the color of deer?

Answer 1.0s

brown

Quantification



Image

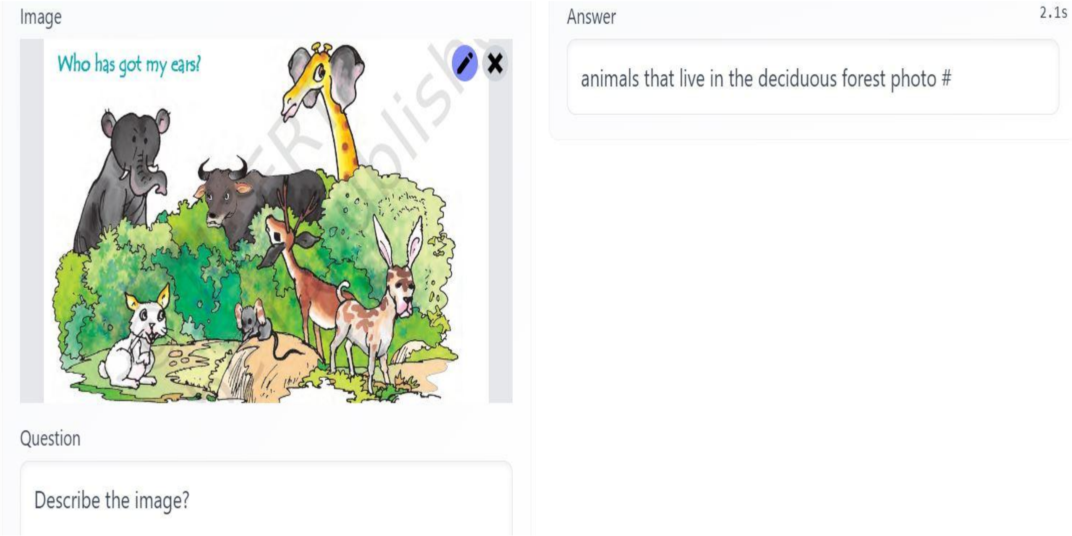
Question

how many animals are present in the picture?

Answer 0.9s

6

Definition



Image

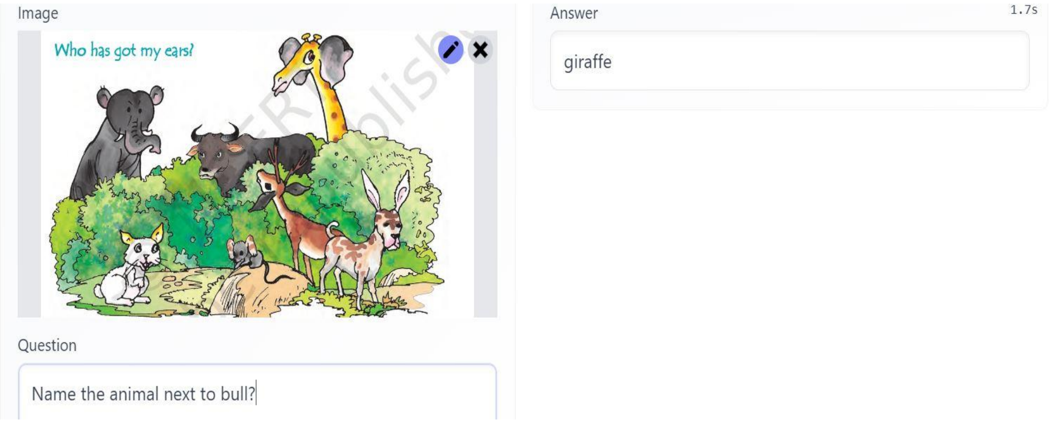
Question

Describe the image?

Answer 2.1s

animals that live in the deciduous forest photo #

One-Word



1.7s

giraffe

Name the animal next to bull?

Sentiment Based

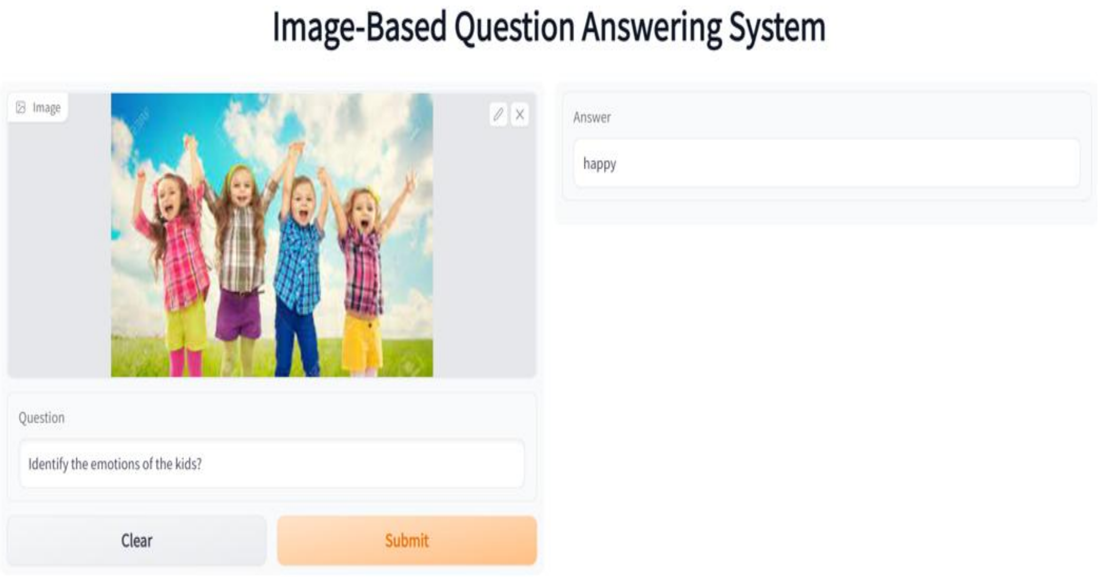


Image-Based Question Answering System

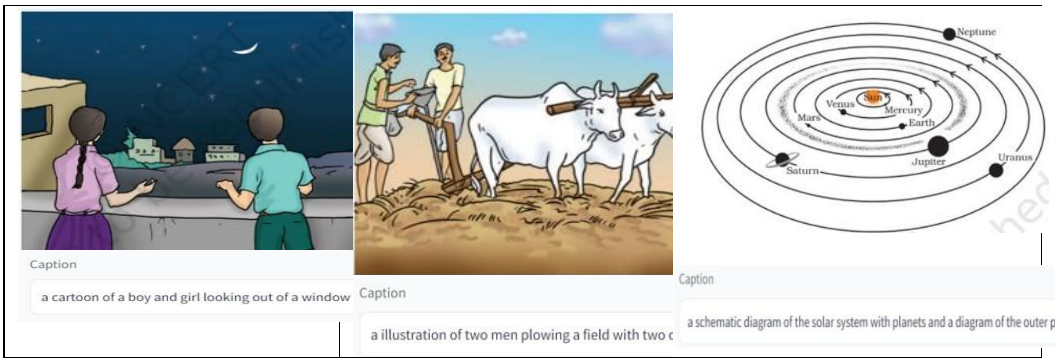
happy

Identify the emotions of the kids?

Fig. 8Testing on different question categories of the system

Generating Description to the Image

For the images where questions were not framed, captions were generated as shown in Fig. 9.



a cartoon of a boy and girl looking out of a window

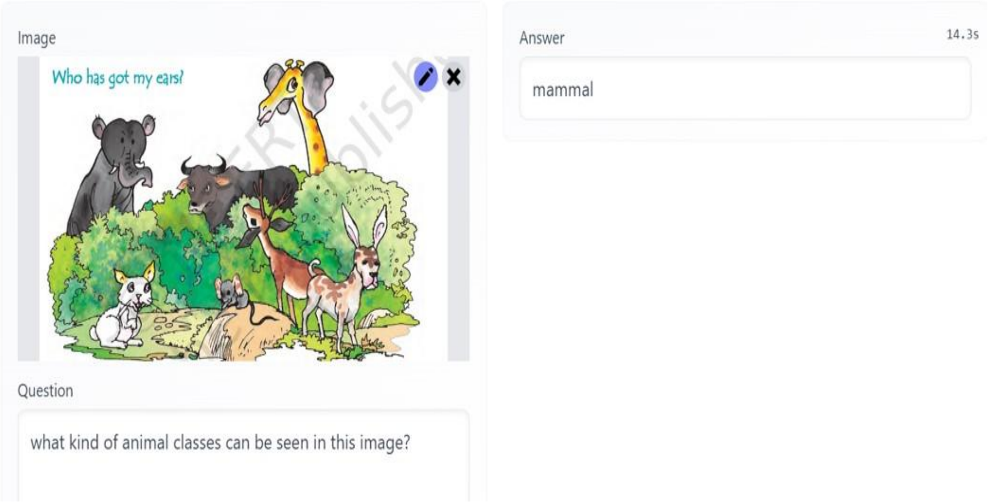
a illustration of two men plowing a field with two c

a schematic diagram of the solar system with planets and a diagram of the outer p

Fig. 9 Testing of EDUIC System on sample images

Knowledge Development

This feature aids in providing better understanding about the image. For example, in the given image system is able to answer the type of animal classes shown in the picture (Fig.10).



Image

Question

what kind of animal classes can be seen in this image?

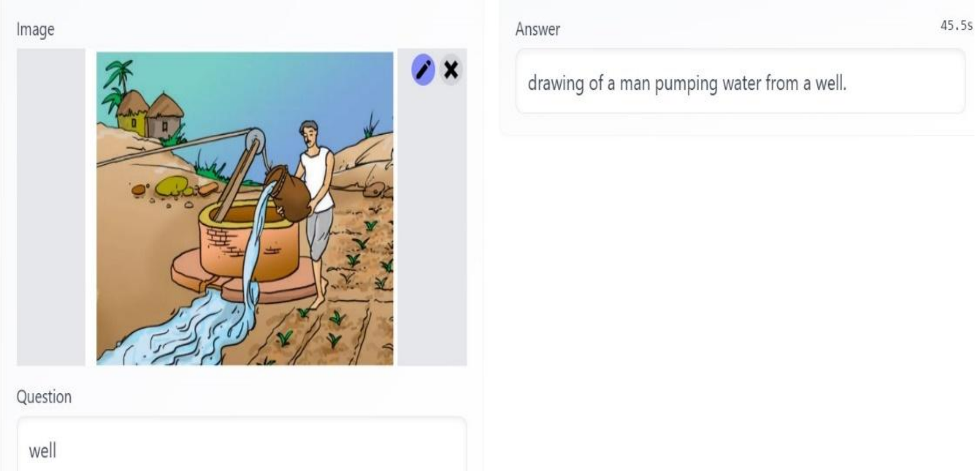
Answer 14.3s

mammal

Fig. 10Knowledge Development

Keyword Based Answer Generation

In this category, an image is taken as an input and keywords related to the image is given as question. The system provides a caption about the image (Fig. 11).



Image

Question

well

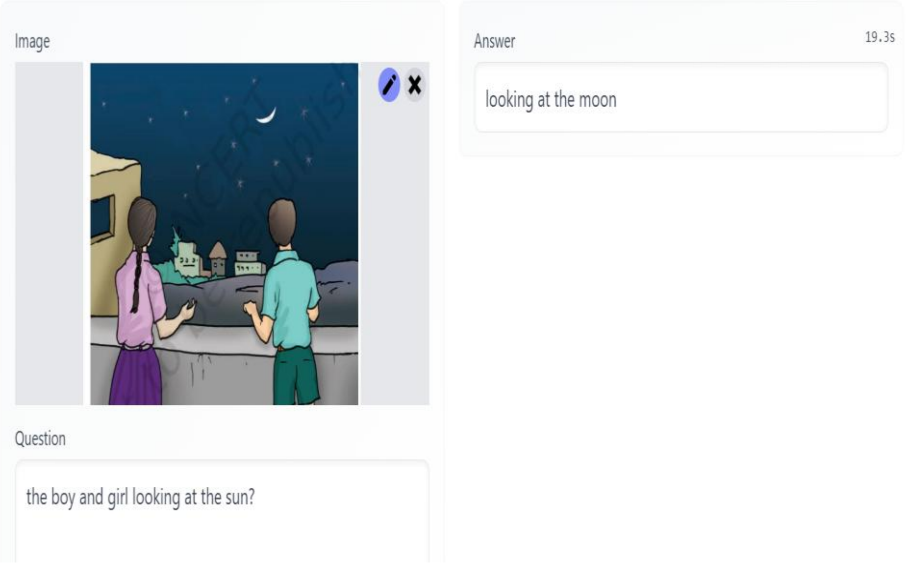
Answer 45.5s

drawing of a man pumping water from a well.

Fig. 11 Keyword based answer generation

Knowledge Correction

Here a wrong related to the image was framed, the system provided the correct answer (Fig. 12).



Image

Question

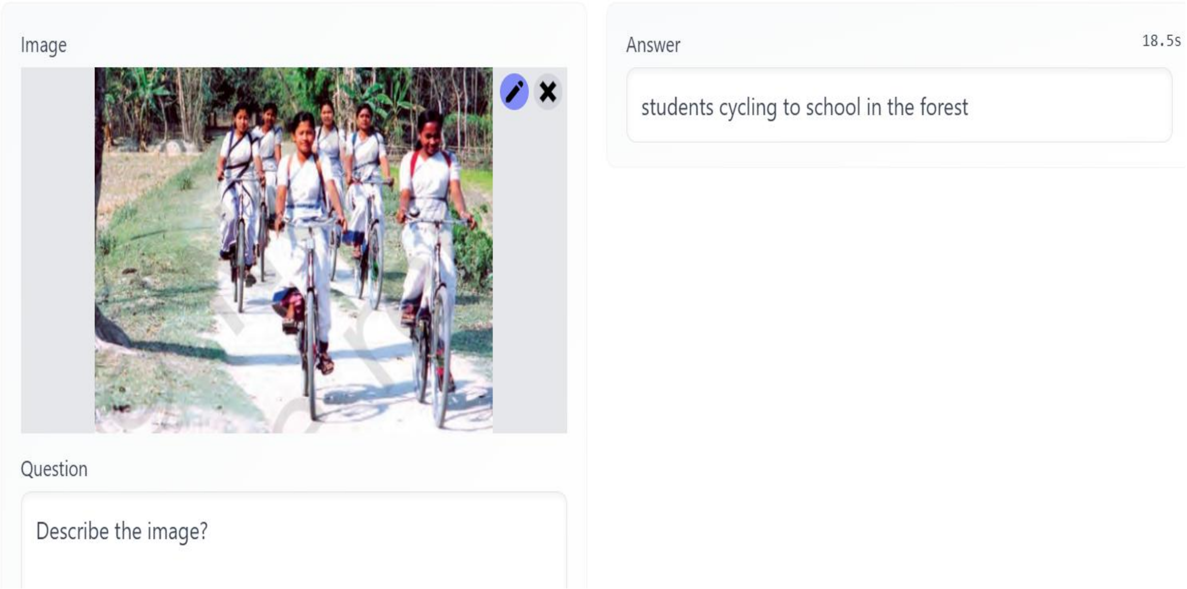
the boy and girl looking at the sun?

Answer 19.3s

looking at the moon

Fig. 12Knowledge Correction

Knowledge of anonymous and complex image for students:
This feature of the system provides the description to an unknown and complex images (Fig. 13).

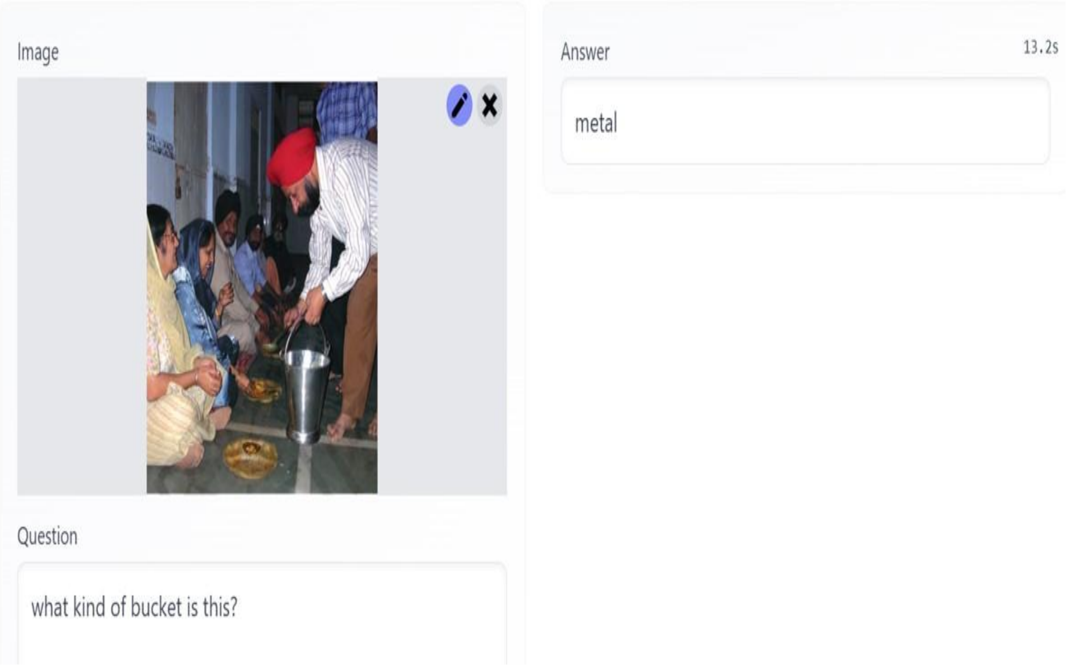


Describe the image?

students cycling to school in the forest

Fig. 13Knowledge of anonymous and complex image for students

Knowledge of arbitrary tested image:
The given image is not stored in EDUVI dataset but the system is still able to answer the question related to the image (Fig. 14).



what kind of bucket is this?

metal

Fig. 14 Knowledge of arbitrary tested image

Evaluation of the System
The developed system was evaluated using various metrics. Firstly, number of correct answers, incorrect answers and number of Not Available (NA) category were counted for every question under each category.It is observed from Fig. 15(a) that the number of correct answer for the category verification based questions is higher as compare to other categories, which revealed that the model is working well on this category. On the other hand, the category quantification based questions contains higher number of incorrect answers.

From Fig. 15(b) it can be analysed that number of correct answers for the category generating descriptions, keyword based and knowledge of arbitrary images are greater as compared to other categories.

Question Categories

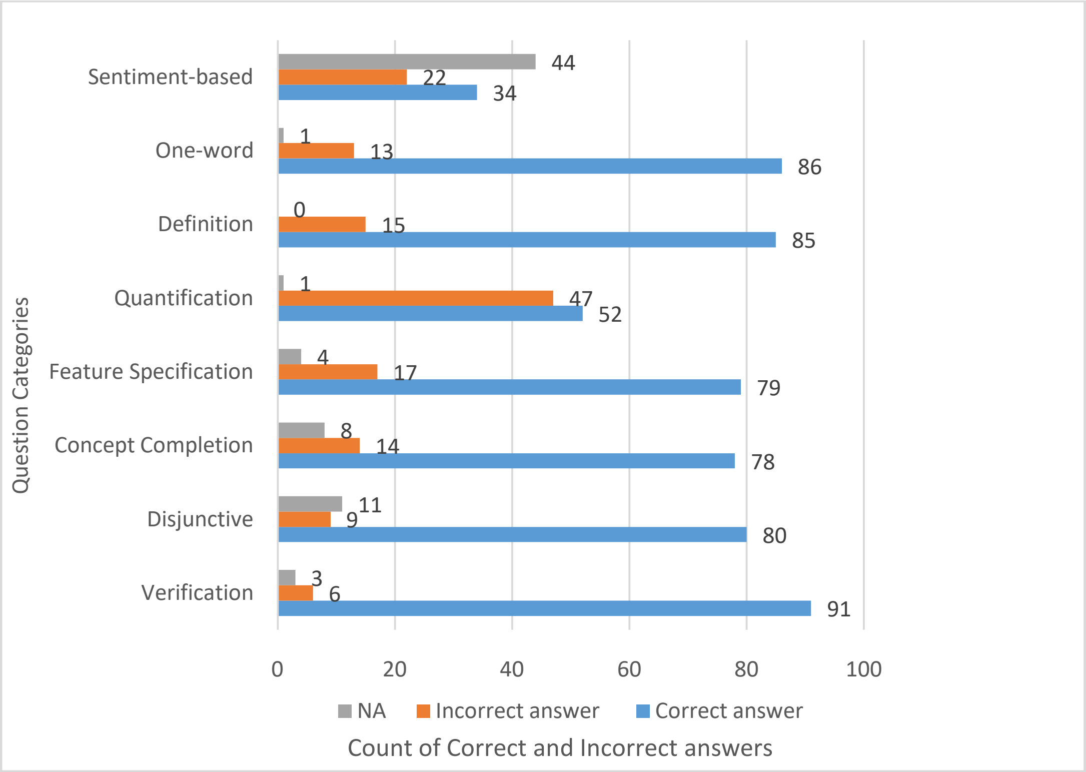


Fig. 15(a) Count of Correct and Incorrect Answers w.r.t. question categories

Knowledge Categories

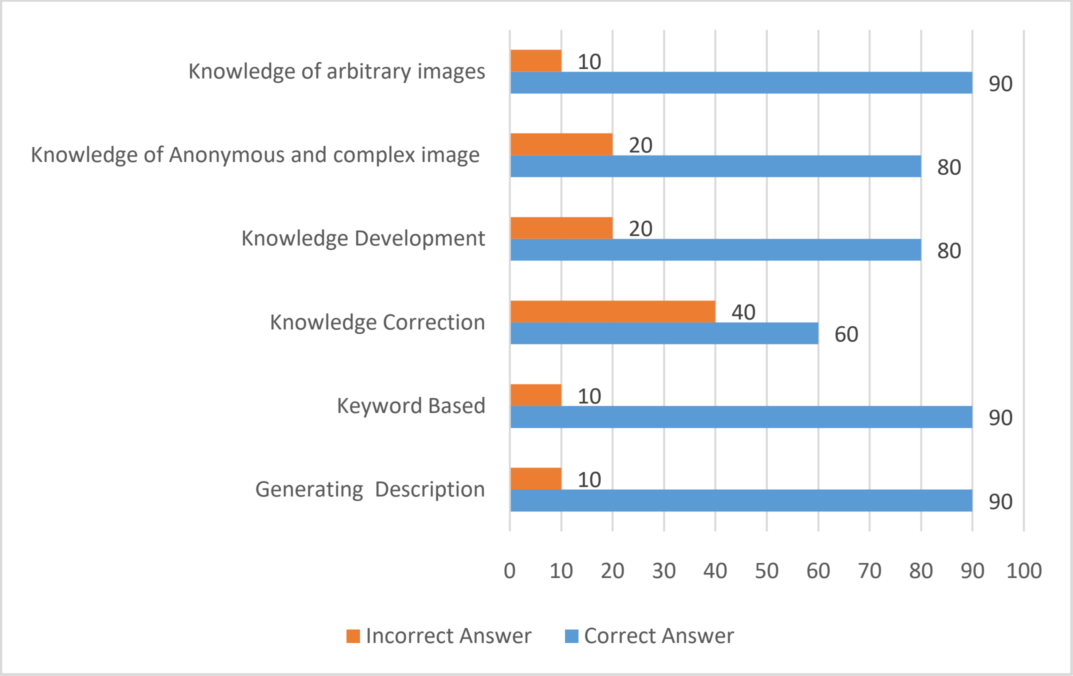
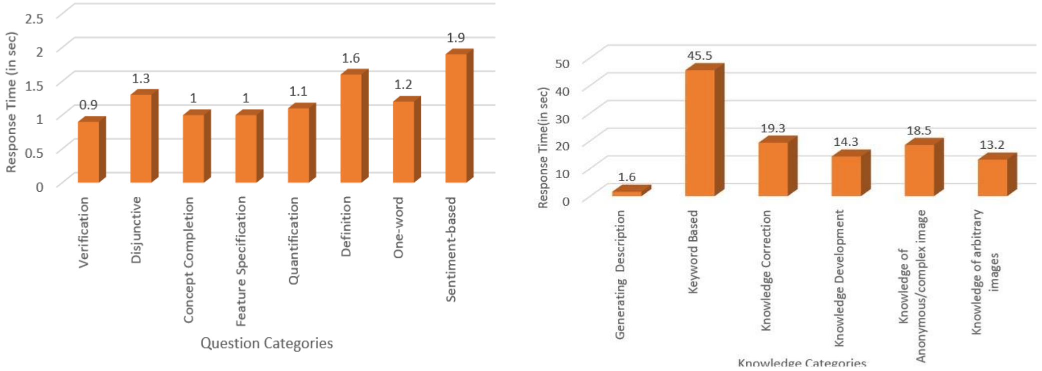


Fig. 15 (b) Count of Correct and Incorrect Answers w.r.t. Knowledge categories

The response time of each question category is shown in Fig.16(a).Itis observed that each type of question category has different response time (in seconds).The response time for the questions in the category sentiment based questions was higher as compared to other categories while the response time for the category verification based question was least.The response time of each knowledge category can be seen in Fig. 16(b) and it can be interpreted from the graph that keyword based category has taken the longest time because it is automatically generating the answerfrom the given keywords related to the image using NLP tasks.

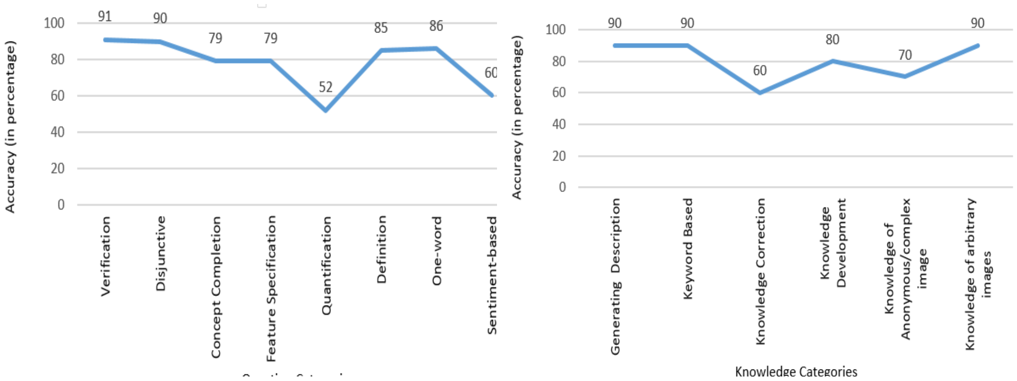


a) Question Categories b) Knowledge Categories

Fig. 16 Response time with respect to (a)Question Categories and(b) Knowledge categories

After testing the model on 800 questions of various categories, accuracy is computed for each category of questions using equation (1). Computed accuracy is shown in Fig. 17(a) and (b).

$$Accuracy = \frac{Total\ no.of\ correct\ predictions}{Total\ no.of\ image\ tested\ on\ particular\ category} \quad (1)$$



(a) Question Categories (b) Knowledge Categories

Fig. 17Comparison of Accuracy on(a) Question Categories(b) Knowledge Categories

From Fig. 17(a) it can be analyzed that the accuracy of the verification based questions is the highest i.e. 91% whereas in Fig. 17(b) it can be seen that generating description, keyword based and knowledge of arbitrary images has the highest accuracy among all the categories.

DISCUSSION

Online learning becomes very popular over the years. Various platforms have been developed to make the teaching and learning style more effective. Visual learning (Jamal Raiyn, 2016) is one of the new approach of learning that enhances the student thinking capability along with learning. Based on the various studies, it can be analysed that students can remember information better when it is represented both visually and verbally. This study develops an EDUVI system to overcome the challenges that are faced by the students during online learning. EDUVI, is an interactive learning platform, which supports visual learning and helps the primary level students in gaining knowledge. The system is proficient of adaptively uniting the outputs from image captioning and visual question answering systems in order to solve a new and more complex problem in education domain. Similar systems have been developed in healthcare (Bansal, Gadgil, Shah&Verma, 2019), tourism (Guerrieri, Ghiani&Manni, 2017) and education domains (Ringe, Marathe, Manjrekar & Shetty,2020) developed a system to educate the stakeholders about how to pronounce words in different languages.The system is novel as it integrates computer vision and natural language processing for education domain which can help the students to solve many doubts by themselves.

For checking the applicability of proposed system, a group of primary section studentswas selected. Analyzing the feedback from the students it was observed that the students appreciated the system. They supported the fact that this system is helpful in enhancing their skill set. They also suggested some improvements which enabled to enhance the functionalities of EDUVI system.

CONCLUSION

This research study mainly emphasizes on developing a visual learning framework, EDUVI, for primary level students that helps them in self-learning without any external assistance. The system consist of two modules:EDUVQA and EDUIC. EDUVI dataset was created by collecting the images from 4th standard E.V.S textbook. CV and NLP techniques are used to process the features of the images and textual features extracted from the question which helps in providing a meaningful answer. The system enables the students in improving the learning and thinking capabilities by answering the questions from different question categories and enhancing the knowledge using different features of the system such as knowledge correction and knowledge development.This work can be extended in future to deal with improving the accuracy of the category quantification based question. Also, the system can be used for understanding the concepts of other subjects for primary to intermediate section. The developed system is capable of answering the questions of images that are related to primary section and can be extended to junior or senior sections by training the model on different type of complex images of various subjects.

REFERENCES

[1]. Al-Sadi, A., Al-Ayyoub, M., Jararweh, Y., &Costen, F. (2021). Visual question answering in the medical domain based on deep learning approaches: A comprehensive study. Pattern Recognition Letters, 150, 57-75.

[2]. Bansal, M., Gadgil, T., Shah, R., & Verma, P. (2019). Medical Visual Question Answering at Image CLEF 2019-VQA Med. In CLEF (Working Notes).

[3]. Brainly (2022) “54% of Indian students comfortable with online learning: Survey”IANS available on https://www.nationalheraldindia.com/national/54-of-indian-students-comfortable-with-online-

[5]. learning-survey. Accessed on 10th June 2022.

[6]. Chebbi, I. (2021). Chabbiimen at VQA-Med 2021: Visual Generation of Relevant Natural Language Questions from Radiology Images for Anomaly Detection. In CLEF (Working Notes) (pp. 1201-1210).

[7]. Cui, Y., Han, G., & Zhu, H. (2022). A Novel Online Teaching Effect Evaluation Model Based on Visual Question Answering. Journal of Internet Technology, 23(1), 91-98.

[8]. Guerrieri, A., Ghiani, G., & Manni, A. (2017, September). A tourist advisor based on a question answering system. In 2017 Intelligent Systems Conference (IntelliSys) (pp. 1173-1176). IEEE.

[9]. Gupta, D., Suman, S., &Ekbal, A. (2021). Hierarchical deep multi-modal network for medical visual question answering. Expert Systems with Applications, 164, 113993.

[10]. He, B., Xia, M., Yu, X., Jian, P., Meng, H., & Chen, Z. (2017, December). An educational robot system of visual question answering for preschoolers. In 2017 2nd International Conference on Robotics and Automation Engineering (ICRAE) (pp. 441-445). IEEE.

[11]. He, X., Zhang, Y., Mou, L., Xing, E., &Xie, P. (2020). Pathvqa: 30000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286.

[12]. Jamal Raiyn(2016) The Role of Visual Learning in Improving Students’ High-Order Thinking Skills Journal of Education and Practice (pp. 115-121)

[13]. Kafle, K., &Kanan, C. (2017). An analysis of visual question answering algorithms. In *Proceedings of the IEEE international conference on computer vision* (pp. 1965-1973).

[14]. Li, Y., Long, S., Yang, Z., Weng, H., Zeng, K., Huang, Z., & Hao, T. (2022). A Bi-level representation learning model for medical visual question answering. Journal of Biomedical Informatics, 134, 104183.

[15]. Liu, B., Zhan, L. M., Xu, L., Ma, L., Yang, Y., & Wu, X. M. (2021, April). Slake: a semantically-labeled knowledge-enhanced dataset for medical visual question answering. In 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) (pp. 1650-1654). IEEE.

[16]. Lu, J., Ye, X., Ren, Y., & Yang, Y. (2022). Good, Better, Best: Textual Distractors Generation for Multiple-Choice Visual Question Answering via Reinforcement Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4921-4930).

[17]. Nugraha, M. H., &Chahyati, D. (2020, October). Tourism object detection around monumennasional (monas) using YOLO and retinanet. In 2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS) (pp. 317-322). IEEE.

[18]. Ringe, S., Marathe, S., Manjrekar, R., & Shetty, R. (2020). Teaching pre-schoolers using VQA: A Web app that answers natural language questions. Zeichen Journal. 6(9), 64-71.

[19]. Siregar, A. H., &Chahyati, D. (2020, October). Visual Question Answering for Monas Tourism Object using Deep Learning. In 2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS) (pp. 381-386). IEEE.

[20]. Toor, A. S., Wechsler, H., & Nappi, M. (2019). Biometric surveillance using visual question answering. Pattern Recognition Letters, 126, 111-118.

[21]. Vo, H. Q., Phung, T. H., & Ly, N. Q. (2020, November). VQASTO: Visual question answering system for action surveillance based on task ontology. In 2020 7th NAFOSTED Conference on Information and Computer Science (NICS) (pp. 273-279). IEEE.

[22]. Wang, T., Li, J., Kong, Z., Liu, X., Snoussi, H., &Lv, H. (2021). Digital twin improved via visual question answering for vision-language interactive mode in human–machine collaboration. Journal of Manufacturing Systems, 58, 261-269.

[23]. Wu, Q., Wang, P., Wang, X., He, X., & Zhu, W. (2022). Medical VQA. In Visual Question Answering (pp. 165-176). Springer, Singapore.

[24]. Yang, L., Cao, H., Hao, F., Zhang, W., & Ahmad, M. (2020, August). Research on tourism question answering system based on xi'an tourism knowledge graph. In Journal of Physics: Conference Series (Vol. 1616, No. 1, p. 012090). IOP Publishing.

[25]. Zhan, L. M., Liu, B., Fan, L., Chen, J., & Wu, X. M. (2020, October). Medical visual question answering via conditional reasoning. In Proceedings of the 28th ACM International Conference on Multimedia (pp. 2345-2354)

[26]. Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., &Gao, J. (2020, April). Unified vision-languagepre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial*

[27]. *Intelligence* (Vol. 34, No. 07, pp. 13041-13049).