International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-11 Issue-1, November, 2021

VisQuelle: Visual Question-based Elementary Learning Companion

A System to Facilitate Learning Word-Object Associations

Sandhya Vidyashankar, Rakshit Vahi, Yash Karkhanis, Gowri Srinivasa

Abstract: ***We present an automated, visual question answering based companion – VisQuelle - to facilitate elementary learning of word-object associations. In particular, we attempt to harness the power of machine learning models for object recognition and the understanding of combined processing of images and text data from visual-question answering to provide variety and nuance in the images associated with letters or words presented to the elementary learner. We incorporate elements such as gamification to motivate the learner by recording scores, errors, etc., to track the learner's progress. Translation is also provided to reinforce word-object associations in the user's native tongue, if the learner is using VisQuelle to learn a second language.***

Keywords: Visual Question Answering; Object Recognition; Question Generation; Question Answering; Word-Object Association.

I. INTRODUCTION

One of the first exposures to reading and writing for children or elementary learners is through alphabet books and charts [1, 2]. These have been posited to help the elementary learners make connections between the symbol (alphabet) on the page or chart with its sound (as taught to the child by a parent or teacher) and the object it represents. For instance, 'A for apple', the symbol 'A' (and sound /ae/ or phoneme /a/) is associated with the fruit [3]. One of the limitations with using a chart or an alphabet book is that the learner may tend to form associations with a particular form of the object.
For instance, Fig. 1 presents a typical image in a letter chart for elementary learning.

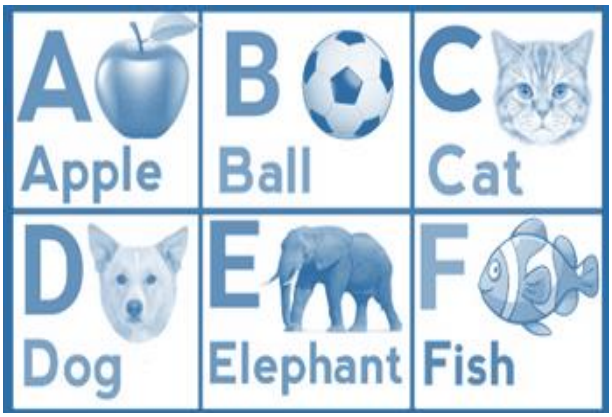


Fig. 1. A snapshot of typical words or pictures associated with an alphabet chart or alphabet book for elementary learning of the English alphabet.

If we take the example of the first word in the chart, 'Apple', a simple search for the fruit 'apple' can lead to one of several possibilities, a few of which are presented in Fig. 2. Further, the word 'apple' can be associated with other words to refer to different fruits such as 'custard apple', 'pineapple', etc. Such interesting, learning opportunities that can be used to introduce a curious elementary learner to adjectives such as numbers (through multiplicity of a noun) and color (varieties of apples, for instance) or increasing the vocabulary or repertoire of nouns (through introducing the names of different fruits or images of different nouns that is spelled starting with an 'A') can be harnessed with a dynamic, visual learning aid.

(a) Red apple (b) Green apple (c) Golden apple

Fig. 2. A sample of apples of different hues – red, green and golden.

Ideally, after the learners have had an initial exposure to the material, the instructor would consider bringing relevant objects to the classroom or, if feasible, take the elementary learners out on a field trip to facilitate 'multisensory learning', i.e., use other senses such as smell, touch and, if feasible, taste, to augment visual and auditory learning [4].

Manuscript received on November 15, 2021.
Revised Manuscript received on November 22, 2021.
Manuscript published on November 30, 2021.
* Correspondence Author
Sandhya Vidyashankar*, PES Center for Pattern Recognition, PES University, Bengaluru, India and Department of Mechanical Engineering, SSN College of Engineering, Chennai (Tamil Nadu), India. Email: vidyashankar.sandhya@gmail.com
Rakshit Vahi, Department of Computer Science and Engineering, PES University, Bengaluru (Karnataka), India. Email: vahi.rakshit@gmail.com
Yash Karkhanis, Department of Computer Science and Engineering, PES University, Bengaluru (Karnataka), India. Email: yashkark@gmail.com
Gowri Srinivasa, Department of Computer Science and Engineering, PES University, Bengaluru (Karnataka), India. Email: gsrinivasa@pes.edu

There is an abundance of literature to support the use of 'manipulatives' to help a learner transition from concrete to abstract concepts [5], as there have been studies on the success of using 'peg blocks' [6] to teach principles of physics, such as 'transfer of energy' or for ways to improve the efficacy of teaching children at the elementary school level to program in a physical programming environment [7]. However, introducing physical aids in a classroom is not always feasible due to economic constraints or the inability of an instructor to access these aids [8].

A. Improvising Teaching Aids for Elementary Learners

There is a compelling case for improvising teaching aids to make learning fun at the elementary level [9]. The pandemic has highlighted the importance of language learning applications with feedback and engagement [10].

Hence, by 'improvisation' we ask, can we design an application that places into the hands of an elementary learner the advantage of time tested association methods and elevate this with an element of 'variety' to stimulate the curious mind and support that a student may have otherwise received from an instructor attending to them?

Further, we seek to include an element of familiarity. As evidenced through the effort of Ranjitsinh Disale, the winner of the Global Teacher Award 2020 that resulted in over 98% of the students achieving the learning outcomes before the end of the school year at a Government School in Solapur, India [11].

Disale took to recording all the lessons of Grades 1-4 in Kannada, the native language of his students and embedded them in QR codes that can be scanned using a mobile phone.

This helped students understand the lessons in their native tongue and relate to the material better, eventually learning what they had to as expected. Thus, the motivation to include a 'translate' feature to help an elementary learner relate to word in their native language as they learn to read and write in a second language or foreign tongue. As we focus on elementary learning, we start with learning to read and write alphabets and focus on word-object associations.

B. How useful are alphabet charts or word books in forming word-object associations?

A study was conducted to evaluate the efficacy of the time-tested approach to introducing children (or first time learners) to reading and writing.

In this study, a group of children were read from an alphabet book for seventeen days and another group of children, the control for this study, were read from a story book during the same time. At the end of this study, children exposed to the alphabet book showed better letter-name recognition than the control group [12]. Another study reports that the forming of letter-name associations translates across languages [13].

To what extent do pictures accompanying the alphabets help children make the 'symbol-object' association? It has been reported that children rarely look at print and the better part of their attention is on the illustration or animation that accompanies the print [14, 15].

II. VISUAL QUESTION ANSWERING (VQA)

A. Question Answering (QA) Systems

Given the importance of an interactive system, we recognize a system that asks questions and is capable of evaluating an answer, would be of value to facilitate the learning process. Computationally, there are multiple Question Answering systems, such as chatbots at the High School and college level [16]. Since, we have established that the target audience comprises children or first-time learners of a language, spoken language recognition systems would have difficulty identifying accents and the learners are not advanced enough to work with complex text interfaces. Hence, the Question Answering System we envisage designing must be capable of accepting questions from users and answering them in a manner that can be understood by the user or to pose questions that can be easily understood and answered by the user.

B. 'Visual' QA

An emerging category of QA systems is visual question answering that involves answering a question that pertains to an image that is presented with the text [17]. The typical vanilla VQA system involves processing an image using a convolutional neural network (CNN) based model, such as VGGNet16 [18], followed by a multilayer perceptron yielding a vector of size, say, 1024 for 1024 image categories and, in parallel, processing an input text (the question) using an embedding (such as GloVe [19]) followed by a sequence-to-sequence model (such as a recurrent neural network (RNN) with a long-short-term-memory (LSTM) [20] followed by a multilayer perceptron that maps the input to another vector of the same size as the image (1024 in our example). These are then combined through a pointwise multiplication, followed by another multilayer perceptron and a softmax that results in probabilities associated with answer choices. See Fig. 3 for a schematic representation of the Vanilla VQA implementation.

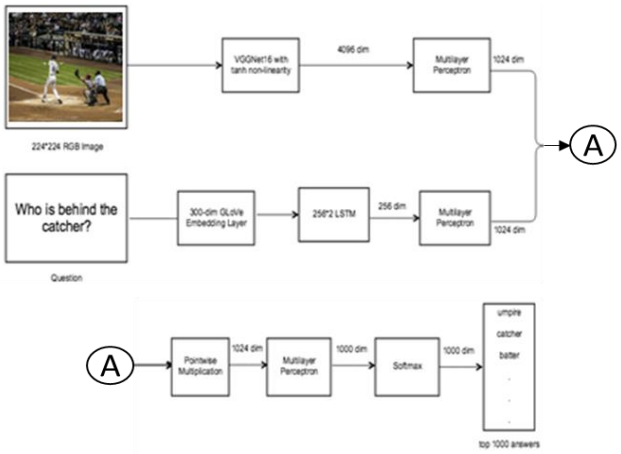


Fig. 3. A schematic diagram of the steps of a vanilla VQA system.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

The first version of the solution proposed in this paper aimed at Gamifying a Vanilla VQA system to work in two stages: In stage 1, the user is required to identify the 'category' of the image (sport versus animal) and in stage 2, the user is required to label a specific instance in the category (type of sport or type of animal) as shown in Fig. 4 [21].

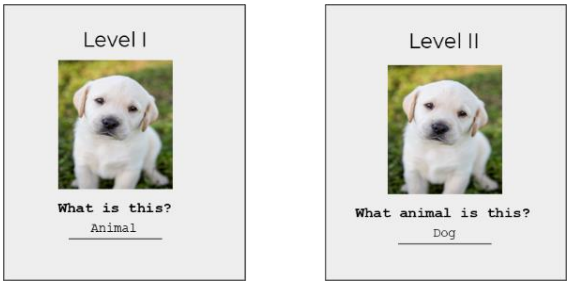


Fig. 4. Two levels of a VQA-based learning aid that identifies a category and then, the specific nature of the input.

It is quite possible with this system that an image is selected that has nothing to do with the question. However, checking for 'visual relevance' – whether the question requires an image to be answered and if the image displayed is an appropriate one for the question was not a part of the workflow. While the Gamification system provides a proof-of-concept of how VQA can be used to aid elementary learning, we seek to enhance the system with the ability to better serve word-object associations and associating with words the learner is already familiar with. In particular, the contributions of the paper are in the capabilities of the Visual Question-based Elementary Learning (VisQuelle) companion that can:

- accept an alphabet or word and go through a graded list of images, presenting multiple forms of the same object, asking the user to locate the object within the image
- ask follow up questions on adjectives – number, color, etc., pertaining to the noun identified
- allow the user to select or upload an image and point to an object that the system will identify and present two alternative images with the word pronounced in the native tongue (applicable for a second language learner whose primary language of preference has been indicated)
- permit the user to ask a question about an object ("What is this object?", "Where is the object in this image?", "How many <nouns> are present in this object?", "What is the color of the <noun> in this object?"etc.) and check for relevance before providing an answer
- accept user feedback from the superuser (course instructor, parent, etc.) to improve the response to the user
- 'gamify' - keep track of scores and a leaderboard, a feature that can be switched on at any point it is necessary to track the progress of the learner or incentivize the learner through a game and score-keeping. The leaderboard is based on the number of questions answered in a set time or the time it takes to answer a set number of questions, in conjunction with the accuracy of the answers.

  The gamification can also be used to encourage a learner to play 'against' a classmate to see who answers the set of questions the fastest.

From a computer vision standpoint, while some of these features have been explored extensively (checking for relevance of the image [22, 23] or asking relevant follow-up questions such as 'how many?' [24], for instance), to the best of our knowledge, the composition of all these elements as a companion to facilitate elementary learning is novel.

The rest of the paper is organized as follows: Section III presents a detailed design of the various components of VisQuelle. Section IV presents experimental results (quantitative measures of performance) for computational models we have trained for VisQuelle and sample images from the relevancy tests and we conclude the discussion in Section V.

III. VISQUELLE: AUTOMATED VQA FOR WORD-OBJECT ASSOCIATIONS

A schematic diagram of the components of VisQuelle are shown in Fig. 5 and expatiated in the sections below.

A. Input

The input is the first component of the system. It can be one of three options: (i) as an alphabet/ word book, the system stores the previous 'page' (alphabet/ word) the user was at and uses that as a trigger to display the image. The user has an option to navigate to the 'next letter' (or 'next word', as applicable), (ii) the user selects from a set of randomly generated images or chooses to upload an image (from the gallery, camera feed, etc.) with a default question on 'What is this picture?' (the default question) or 'Is there a (or an) <adjective/ noun>?' (to be keyed in by the user) or an easier alternative to the previous option, viz., 'A word that starts with <letter>' (gap fill – letter to be keyed in by user).

B. Processing – Natural Language Process (NLP)

In the first step, if any letter is input by the user, this is extracted and used to match tags associated with images in the database for the selection of an appropriate image, if this step is necessitated by the input.

If a word (or adjective, etc.) is input by the user, this is matched with the image tags in the database as well as a list of synonyms to find image tags that are a likely match and used to generate an image.

If a question is input (or generated) with the randomly selected image, then, rather than process the words of the question literally, we find the parts-of-speech (PoS) tags for the words. PoS tags capture the structure of the sentence (for visual relevance) better than the actual words themselves and generalize well across image categories.

C. Processing – Object Detection

We use a RetinaNet model trained on the Common Objects in Context (or COCO) dataset to perform object detection of an input image [25]. The RetinaNet model uses multiscale features to augment a traditional

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

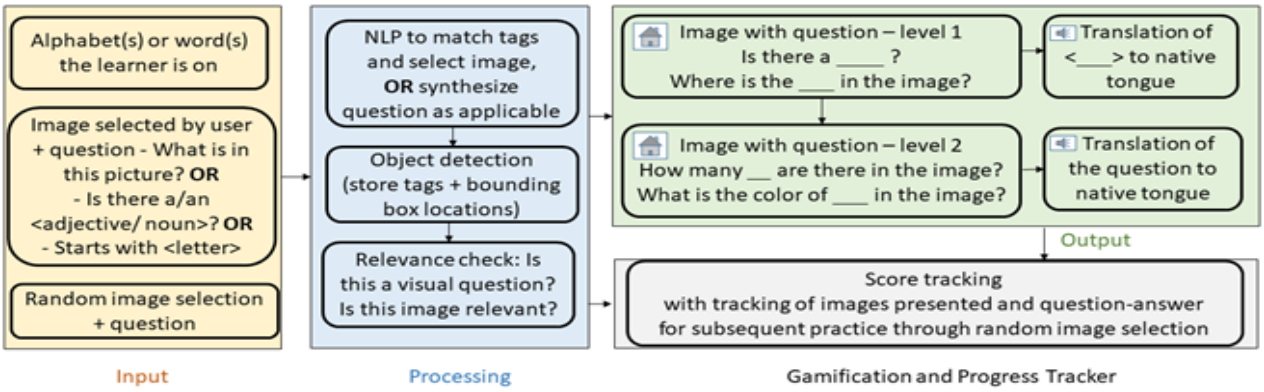


Fig. 5. Schematic diagram of the components of VisQuelle

feedforward neural network with the ResNet architecture. The output is a bounding box that is centered at each object detected. With over 2.5 million images in the COCO dataset, RetinaNet is trained to recognize 80 common images and is amenable to generalization [26]. The advantage of this model is that we have a bounding box for every instance of an object as well as bounding boxes for most common categories of objects. These are saved for use the next step of the processing, viz., relevance matching.

D. Processing – Relevance Check

If a question is input with an image, the relevance check is applied to the image-question pair. This is done in two stages (see Fig. 6 for a schematic diagram of the steps).

Visual Relevance: Initially, we check whether the question pertains to an image or not. POS tags of the words in the question (see Section III. B.) are used to train an LSTM network in conjunction with the corresponding images and a label that is 'true' for questions that require an image to arrive at an answer and 'false' for questions that are independent of an image (such as 'Who is the President of <country-name>?'). The images and questions are selected from the same dataset, with annotated images and questions being presented as 'positive' samples and a combination of the images with questions tags associated with images that have no overlap in the objects (based on the low or no similarity between question tags) as 'negative' samples.

Relevance of the input image: If the question requires an image to arrive at the answer, then, it is tagged 'visually relevant' and object detection (as described in Section III B) is applied using RetinaNet. The output of this step is bounding boxes around every instance of an object that is recognized in the image, with tags for each bounding box, describing the object detected. The input question is processed in parallel to remove stop words or to extract key words (essentially, nouns, adjectives, etc.) in the question.

Next, a list of synonyms is generated for the tags output by the RetinaNet and those output from the question after the removal of stop words. This ensures, while nouns may not be referred to in exactly the same manner (for instance, 'car' and 'auto (mobile)' could well refer to the same object in the image), their similarity is still accounted for. Finally, an intersection is performed between the two processed text lists (tags generated from RetinaNet with its synonyms and key words from the question with its synonyms). If the question is relevant to the image, the intersection is expected to be nonempty. If the intersection is a null set, then the question is declared to be one that requires an image but that the input image is not the relevant one to answer the question (or, if the image is input by the user, then the question is declared to be not relevant to the input image).

As depicted in Fig. 6, if we find the question asked does not require an image to output an answer, we find the answer using Google API and IBM Watson [27], using the agreement between the two as a confirmation of the 'answer'. This is not currently output as a part of the VisQuelle Companion.

E. Output

Once we have determined the relevance of the question to the image (or vice-versa), the output follows one of multiple formats.

VisQuelle as an alphabet book with 'unlimited' pictures: If the user intends to use the VisQuelle Companion as an electronic equivalent of a book, the QA component is switched off and only the image is presented with tags. Toggle buttons on the user interface permit a user to present a new image for which the tag remains the same or to proceed to the next letter in the alphabet (or next word in the list, etc., depending on the grade/ progress demonstrated by the learner).

Since pictures are considered a source of 'distraction' for an elementary learner [14, 15], we limit the number of pictures to 10 by default, but the number can be set to any number that a guardian or instructor finds suitable.

The images are sampled with replacement; this means, an image can be presented more than once and would serve to reinforce a word-object association.

VisQuelle as a yardstick for object-recognition: If the interactive mode is switched on (as it is, by default) then, the system displays the image and presents a level-1 question, asking the learner to 'point' to the object of interest (typically, sampled from the tag) describing objects in the image. The (x,y) coordinates of the cursor, controlled by hand, a mouse or stylus is recorded. A left click on the mouse (or a touch of the screen), etc., within the bounding box of the object mentioned in the question is considered the right location. After three tries, in case the user fails to locate the object,

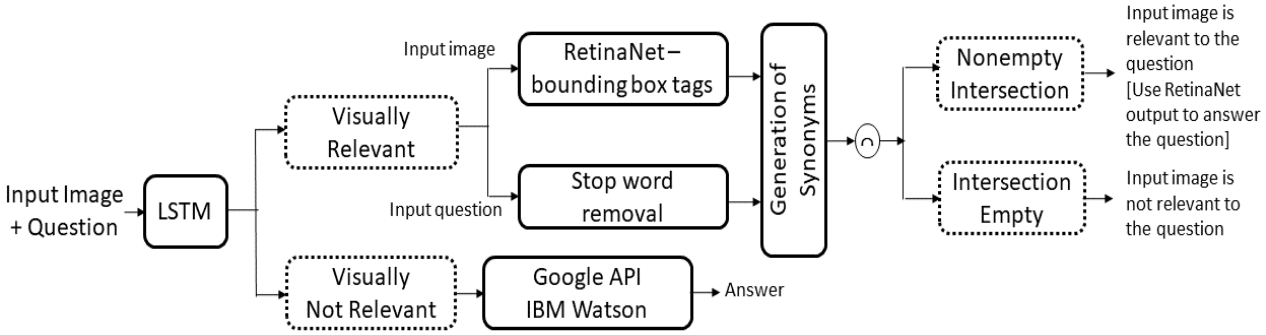


Fig. 6. Relevance check in two-steps: (i) is the question visual? and (ii) is the input relevant to the image?

VisQuelle presents a red bounding box around the object concerned and prompts the user to try with another image for the same input (letter or word).

VisQuelle as an elucidator: A level-2 question comprises asking a follow-up question on image – this pertains to a description that can be assessed using image tags from RetinaNet post the object recognition phase. Typical questions pertain to number ("how many…") and the answer is based on the number of bounding boxes associated with a tag that matches the word or its synonym.

Another question pertains to the color. This is done through an analysis of the 'dominant color' in the image. Hexa-decimal color codes in the Red, Blue and Green (or the RGB) color space are quantized and mapped to one of the colors present in a list comprising common colors names ('red', 'blue', 'green', etc.). The palette can be made more specific with a finer quantization to describe different shades of a color, etc. However, considering our user-base comprises elementary learners, we have limited the list.

VisQuelle as a translator: For learners for whom English is a second language, VisQuelle attempts to harness word-object associations that may have already been formed in another language. Through recall and repetition, the association is reinforced in English. This is done using Google's Translate API and a simple phoneme-stitching based synthesis of speech. This can replaced with more meaningful recordings of words in multiple native tongues to better benefit underserved communities, where learners have a higher dependence on a computational support such as VisQuelle.

F. Score Tracking

VisQuelle can be used to keep track of the progress of a user through a login. Their profile would be updated with each use of the system, recording word-object associations the user has identified correctly against those the user needs more practice with. These are used to present suitable examples in the Gamification ("Quiz") phase, where points are associated with each answer. If an instructor finds it appropriate, they can enable a leaderboard for their class to present top scorers. The leaderboard feature is switched off by default to avoid causing stress to elementary learners.

G. Role of VQA in VisQuelle and Feedback

When a random image is input to the system with a question, what is the most likely answer?

This is computed using the Vanilla VQA system from the predecessor model presented in Fig. 3.

This uses a VGG-16 model, where 16 refers to the number of convolutional neural networks (CNN). Since the final output is based on probabilities, it is possible that the system may have some faulty tags.

To this end, there is a review of word-object associations (i.e., automatically generated image-annotations) that can be manually corrected through the feedback of a course instructor or guardian. If the system is not a single-user one, the proportion of 'wrong' tags generated for an image can also be used to automate the filtering of images/ objects that need a review.

IV. EXPERIMENTAL RESULTS

We have performed unit testing of each component that we trained afresh and tested for the purpose of Elementary Learning.

The first of these is the visual relevance model. From Table 1, we note that precision is high for both categories, while recall is high for samples where the question is irrelevant to the image.

The consequence of this is that we may categorize something as 'irrelevant', even when it is 'relevant'.

This prompts us to run the level-2 test with RetinaNet anyway, as a confirmation of the relevance of the question to the image. Sample images from the relevance tests, stage 1 (corresponding to Table I) are presented in Fig. 7.

We note that the user is prompted to point to an object within the image. The coordinates are recorded.

Table- I. Visual relevance

SYMBOL	RECALL	PRECISION
RELEVANT	0.732	0.956
IRRELEVANT	0.997	0.978

As it happens in the first row of Fig. 7, if the coordinates are located with the bounding box for the object, the system recognizes that as a correct response. As it happens with the second row of images, if the coordinates lie outside the bounding box, then the user is prompted to try again. After three unsuccessful attempts, the system displays the bounding box around every instance of the object in the image.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Fig. 7. Relevance check stage-1 in two-steps: (i) is the question visual? and (ii) if yes, is the input location correct?

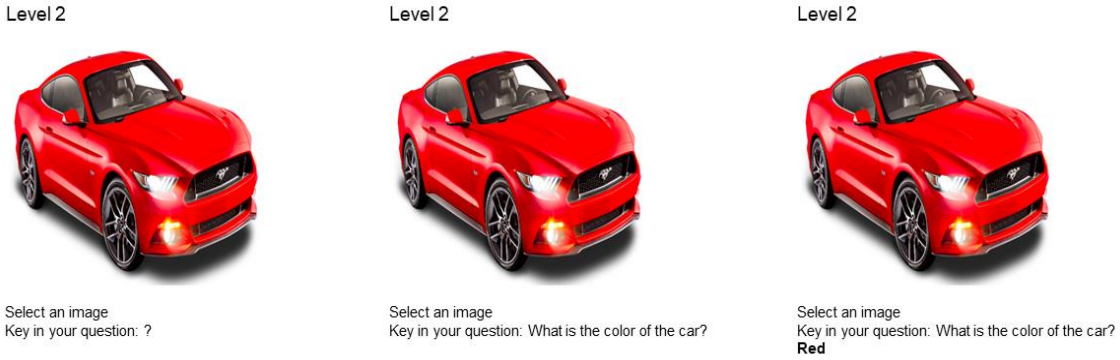


Fig. 8. Relevance check: stage-2: the user is asked to select an image and key in a question. If the question is visual (stage 1) and relevant to the image (stage 2), a response is returned on the basis of the tags from Retina Net and mapping of dominant colors.

Table- II. Relevance of the question to the input image

SYMBOL	RECALL	PRECISION
RELEVANT	0.851	0.887
IRRELEVANT	0.971	0.961

From Table II, we note that both recall and precision decrease in the case of the level-2 relevance testing.

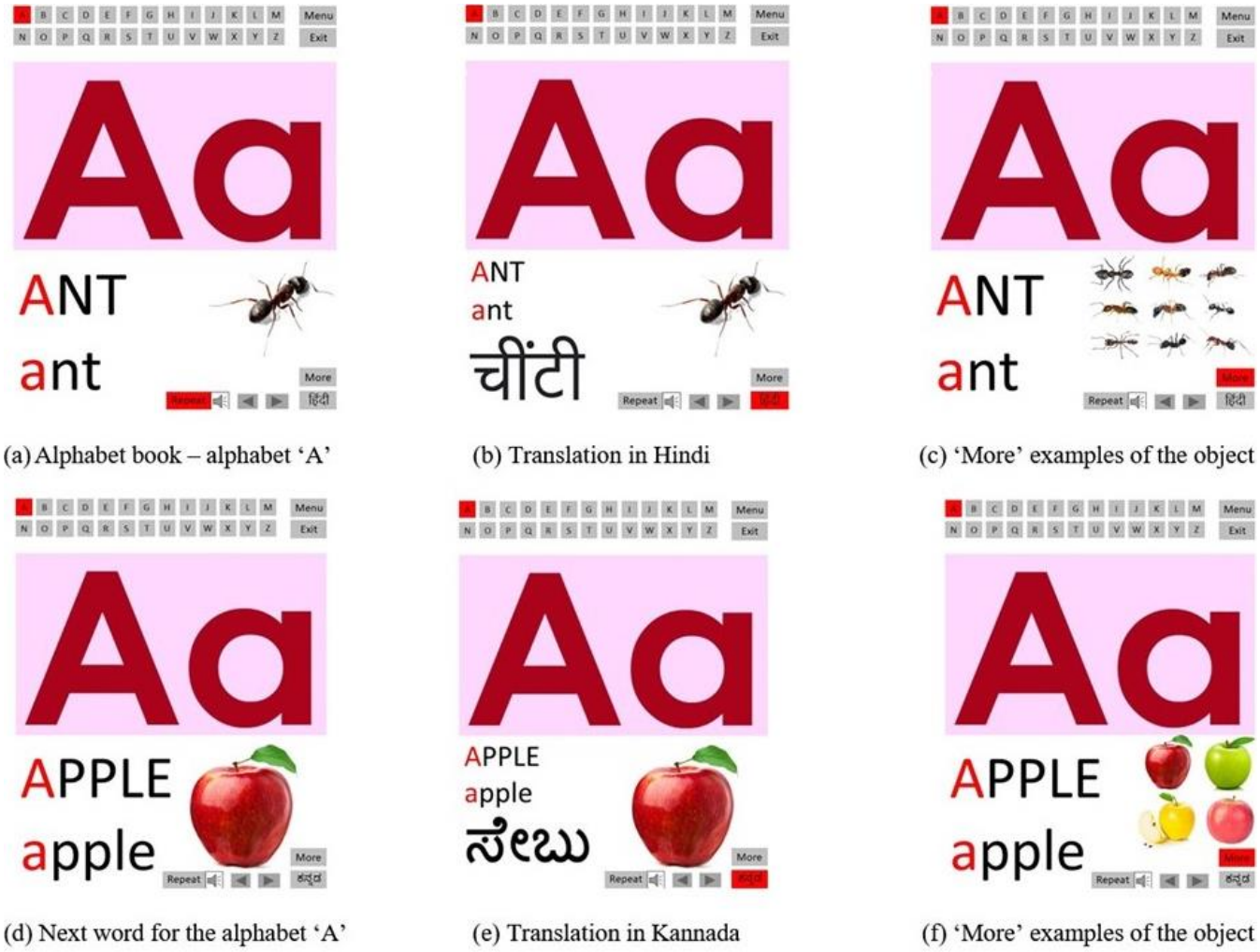
The irrelevant class still shows higher recall and precision than the relevant class. When compared to the level-1 tests, recall of relevant images is higher. We think the extent of detail ensures a better categorization of the categories. Hence, the decision to run level-2 test as a confirmation even for level-1. Sample images from the relevance tests, stage 2 (corresponding to Table II) are presented in Fig. 8.

This is an example where a user has selected an image and is prompted to key in a question. For the example shown in Fig. 8, the probabilities for the top five answers are presented in

Table- III. Probabilities for the top five results from VQA corresponding to Fig. 7

TOP FIVE RESULTS	PROBABILITY
RED	0.7245
RED AND WHITE	0.0854
BLACK	0.0669
ORANGE	0.0344
WHITE	0.0275

We note that the probabilities returned by the VQA model ‘as is’ are based on the proportion of occurrence of the colors in the tags, rather than an actual analysis of the content of the image. The top five options account for about 0.94 of the probability, of which the correct option happens to be the top choice in this example. To circumvent this uncertainty, we prefer to use the mapping to dominant colors as a primary approach to solving this problem and to use the output from VQA as a check.



(a) Alphabet book – alphabet 'A' (b) Translation in Hindi (c) 'More' examples of the object

(d) Next word for the alphabet 'A' (e) Translation in Kannada (f) 'More' examples of the object

Fig. 9. VisQuelle as an alphabet book (a) a word and its image (b) with translation of 'ant' in Hindi (c) 'more' examples for ant, (d) the 'next word' in the alphabet book – 'apple'; the primary language has also been changed (e) 'apple' in Kannada and (f) 'more' examples for apple.

Table- IV. System testing

TYPE OF QUESTION	ACCURACY (%)
DOMINANT COLOR	76
YES/ NO (PRESENCE/ ABSENCE) FOR AN OBJECT	52
HOW MANY	20

The accuracy at the system level for various question-types is presented in Table III. We note that dominant color is the highest with 76% whereas even yes/ no questions and counting type questions perform poorly. These issues are mitigated to a large extent when explicit processing of the text is done in conjunction with the processing of the images as shown in Fig. 6. Performance measures for features of the VisQuelle system derived from pre-trained or open source models available in the public domain (such as the LSTM, RetinaNet, etc.,) are reported in the relevant papers. Other components (such as using VisQuelle as an alphabet book) are deterministic functions, much akin to a database retrieval.

VisQuelle as an alphabet book: A couple of examples of the alphabet book feature are presented in Fig. 9. It is seen that the options within the 'alphabet book' mode usage of VisQuelle allows a user to simply view the letter with the word and an image of the object – this is akin to a chart or picture book for word-object associations or to learn the letters of the alphabet as shown in Fig. 9(a) and 9(d). The option 'repeat' allows the user to listen to the alphabet and word again (any number of times). The second option, highlighted in red in Fig. 9(b) and in Fig. 9(e) presents a translation in the learner's native tongue to reinforce concepts and build on previously learned associations. The options available are (i) Hindi, (ii) Kannada and (iii) Tamil and can be extended to other languages. The third option allows the user to see 'more' examples of the object instance (see Fig. 9(c) and Fig. 9(f)). The accuracy of these subcomponents is almost 100% as it depends on the accuracy of the annotations provided with the images that are retrieved and these, being manually curated, can be expected to be completely reliable. We can incorporate more content based retrieval techniques for image and text [28, 29] to enhance the ability of current system, however, for an elementary learner, these do seem to suffice. The vocabulary is adopted from charts available and through studying multiple curricula offered in schools. Alternative object associations for 'A' include, besides apple, 'ant', 'animal', 'angry' (taught with 'emotions'), 'above'

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

(taught with 'directions'), 'Aunty' (taught with 'family'), 'airplane' (taught with 'vehicles') and 'army' (repeated with collective nouns at a higher grade – an 'army of ants') etc. If it be asked, 'Why not curate a dataset for retrieval?' the answer to this is in the motivation – we would like elementary learners to have an opportunity to explore nuances of object instances, even as they learn word-object associations, so the effort is not akin to learning by rote.

Moreover, these systems can be adapted for focused learning, such as 'class of vegetables' or 'class of animals', etc. Despite these advantages, we are unable to design a solution that is entirely free from constraints for the reason 'images in the wild' may actually confuse a learner who is depending to a large extent, if not entirely, on a system like VisQuelle to support their learning. For instance, one of the first few searches on Google Images for the word 'Apple' includes a logo of the company by the same name. While we seek to add some nuance to the input, such drastic variations from the 'theme' may be counterproductive.

V. CONCLUSION

In summary, we motivated the need for an interactive learning aid to support elementary learning and formation of word-object associations. With the backdrop of the Covid-19 pandemic, a digital learning companion is a valuable resource. In particular, the addition of nuance and an exposure to various instances of an object, is a value add we get from a visual-question based system that supports learning word-object associations. Additionally, gamification and score-tracking help an instructor or guardian track the progress of a learner and may reveal patterns indicative of difficulties, if any, through the learning process. Incorporation of a translation module to map words to the native tongue of the user helps to reinforce word-object associations in a second language/ foreign tongue. VisQuelle is a proof-of-concept system that addresses all these needs.The accuracy of the underlying machine learning models for object recognition, etc., trained on subsets of unconstrained categories of images, is limited. We overcome this in the context of our application by limiting the vocabulary to words associated with those typically taught at the elementary level. The performance can be improved through more focused datasets and better trained models for object detection and recognition and more robust processing of images and associated input text.

REFERENCES

1. P. Nodelman, 2001 A is for...what? The function of alphabet books. Journal of Early Childhood Literacy, vol. 1, pp. 235--253.
2. K. S. Bradley, J. Bradley, 2014 Using Alphabet Books across Grade Levels: More than 26 Opportunities. Texas Journal of Literacy Education, vol. 2(1), pp. 13-23.
3. A. Castles, M. Coltheart, K. Wilson, J. Valpied, J. Wedgwood, 2009 The genesis of reading ability: What helps children learn letter–sound correspondences? Journal of experimental child psychology 2009, vol. 104, no. 1, pp. 68—88.
4. L. Blomert, D. Froyen, 2010 Multi-sensory learning and learning to read. International Journal of Psychophysiology, vol. 77(3), pp. 195– 204. https://doi.org/10.1016/j.ijpsycho.2010.06.025
5. M. Boggan, S. Harper, and A. Whitmire, 2010 Using Manipulatives to Teach Elementary Mathematics. Journal of Instructional Pedagogies, pp. 3.
6. B. Piper, I. Hiroshi, 2002 PegBlocks: a learning aid for the elementary classroom. CHI'02 extended abstracts on Human Factors in Computing Systems.
7. K. H. Jin, H. Kathleen, K. Gavin 2016 Teaching elementary students programming in a physical computing classroom. Proceedings of the 17th annual conference on information technology education.
8. O. T. Ibeneme, 2000 Provision and utilization of instructional equipment for teaching and learning science and technology. Issues in Educational Journal, vol. 1, pp. 139-144.
9. M. O. Asokhia, Improvisation/teaching aids: Aid to effective teaching of English language. 2009 International Journal of Educational Sciences, 2009, vol. 1(2), pp. 79-85.
10. J. Egbert, 2020 "The new normal?: A pandemic of task engagement in language learning. Foreign language annals, vol. 53(2) pp. 314-319.
11. Global Teacher Prize, "Ranjitsinh Disale", 2020: https://www.globalteacherprize.org/person?id=13756 (Last Accessed on 31st August, 2021.)
12. A. C. Both-De Vries, and A. G. Bus, 2014 Visual processing of pictures and letters in alphabet books and the implications for letter learnng", Contemporary Educational Psychology, vol. 39, pp. 156--163. http://dx.doi.org/10.1016/j.cedpsych.2014.03.005.
13. R. Treiman, I. Levin, B. Kessler, 2007 Learning of letter names follows similar principles across languages: Evidence from Hebrew. Journal of Experimental Child Psychology, 96(2), 87-106.
14. L. M. Justice, P. C. Pullen, K. Pence, 2008 Influence of verbal and nonverbal references to print on preschoolers' visual attention to print during storybook reading. Developmental Psychology, vol. 44, 2008, pp. 855—866. http://dx.doi.org/10.1037/0012-1649.44.3.855.
15. M. A. Evans, J. Saint-Aubin., N. Landry, 2009 Letter names and alphabet book reading by senior kindergarteners: an eye-movement study", Child Development, vol. 80, pp. 1824—1841. http://dx.doi.org/10.1111/j.1467-8624.2009.01370.x
16. J. Ureta, P. J. Rivera, 2018 Using Chatbots to Teach STEM Related Research Concepts to High School Students. Workshop Proceedings, International Conference on Computers in Education, Philippines: Asia-Pacific Society for Computers in Education, pp. 338.
17. S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, L. C. Zitnick, D. Parikh, 2015 VQA: Visual question answering. In: IEEE ICCV, pp. 2425–2433.
18. K. Simonyan, A. Zisserman, 2014 Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 2014.
19. J. Pennington, R. Socher, C. D. Manning, 2014 Glove: Global vectors for word representation, Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543.
20. S. Hochreiter, J. Schmidhuber, 1997 Long short-term memory. Neural computation, vol. 9(8), pp. 1735–1780.
21. S. Suresh, V. N. Rao, G. Srinivasa, 2018 Gamification of a visual question answer system. Proc., IEEE tenth international conference on technology for education.
22. P. Prabhakar, N. Kulkarni, L. Zhang, 2018 Question Relevance in Visual Question Answering. arXiv preprint, arXiv:1807.08435.
23. Y. Goyal, 2017 Making the V in VQA matter: Elevating the role of image understanding in visual question answering. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
24. A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. F. Moura, D. Parikh, D. Batra, 2017 Visual dialog. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 326-335.
25. T. Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, "Focal loss for dense object detection" In Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980-2988.
26. T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona D. Ramanan, C. L. Zitnick, 2014 Microsoft Coco: Common objects in context", European conference on computer vision, pp. 740-755. Springer, Cham.
27. E. Davis, M. Gary, 2015 Commonsense reasoning and commonsense knowledge in artificial intelligence 2015 Communications of the ACM, vol. 58(9) pp. 92-103.
28. K. Shriwas, V. Ansari, 2016 Content based Image Retrieval using Model Approach. International Journal of Applied Information Systems 10(8):27-32.
29. F. T. Da Silva, J. E. B. Maia, 2020 Luppar: Information Retrieval for Closed Text. International Journal of Applied Information Systems 12(28).

AUTHORS PROFILE

Sandhya Vidyashankar is a Mechanical Engineering Undergraduate student at SSN College of Engineering, Chennai, India, and has completed an internship at the PES Centre for Pattern Recognition. Her research interests lie in Mobile Robotics, Computer Vision and Intelligent Systems.

Rakshit Vahi is a Computer Science and Engineering graduate from PES Institute of Technology, Bangalore South Campus (2019), Bangalore. His research interest lies in machine learning. He wants to pursue his career in the healthcare domain.

Yash Karkhanis is a Computer Science and Engineering graduate from PES Institute of Technology, Bangalore. He is currently working as a Senior Software Engineer in a company which builds AI based software products for the Healthcare industry. His main areas of interest are NLP and computer vision and how they are applied in the healthcare industry.

Gowri Srinivasa obtained her PhD in Biomedical Engineering from Carnegie Mellon University, Pittsburgh, USA (2008). Since August 2008, she has been with PES University, Bengaluru, India, where she is currently a Professor in the Department of Computer Science and Engineering and heads the Center for Pattern Recognition. Her research focuses on the design of signal processing, analytics and machine learning based solutions to problems primarily in the domains of education and healthcare. She is a Senior Member of the IEEE and has served on the Technical Committee (TC) of the IEEE Bio Imaging and Signal Processing Society (BISP) during 2009-2019 and on the TC of the IEEE Conference on Technology for Education (T4E) since January 2019.