Experiential questioning for VQA

Ruben Gómez Blanco, Adrián Pérez Peinador, Adrián Sanjuan Espejo, Antonio A. Sánchez-Ruiz[0000−0003−0034−1254] and Belén Díaz-Agudo[0000−0003−2818−027X]

Department of Software Engineering and Artificial Intelligence
Instituto de Tecnologías del Conocimiento
Universidad Complutense de Madrid, Spain
{rubgom03,adpere08,adrisanj,antsanch,belend}@ucm.es

Abstract. Visual Question Answering (VQA) is a task born out of the need to answer queries regarding images or videos. Unlike simpler tasks such as classification or regression, VQA requires expertise from both computer vision and language modeling domains. These systems typically mimic human reasoning by detecting objects and establishing their relationships within the image using different techniques such as object detection, fine-grained recognition, action detection, and common-sense reasoning. VQA systems generally assume that the user initiates the interaction by asking specific questions about the image, but this can be problematic for some people with visual impairments. In this paper, we present a case-based approach to help users formulate relevant questions about an image based on questions that other users asked about similar images. We evaluate the use of different similarity measures between images and propose a way to cluster and filter the retrieved questions.

Keywords: Case-Based Reasoning, Visual Question Answering, Similarity

1 Introduction

Visual Question Answering (VQA) is a task that emerged from the necessity of answering questions about an image or video. Unlike simpler tasks such as classification or regression, VQA requires expertise from two key domains: computer vision and language modeling. It seems in general VQA systems do follow human reasoning in the sense that they detect objects in the image and then establish the relations between those objects [3]. Research in VQA has identified different types of questions regarding the knowledge needed to answer them [2]: object detection, like *Are there dogs in the picture?*, fine-grained recognition, like *What type of dog breed appears in the picture?*, action or activity detection and recognition, like *Is the dog eating?*, others knowledge-intensive or common-sense questions, like *Is the dog breed the favorite dog breed of Queen Elizabeth II?* or *Does the dog love her humans?* [4].

VQA is also useful in helping the visually impaired describe or understand the content of a photo and is invaluable in fostering independence and enriching experiences by allowing users to interact with visual content independently [14]. However, most VQA systems tend to assume that it is the user who initiates the interaction by asking specific questions about the image. This assumption can be problematic for some people, particularly those with visual impairments, who, unable to see the image, can only ask very general questions and may need several iterations of questions and answers to get the relevant information from the image.

In this paper, we present a case-based approach to help users formulate relevant questions about an image based on questions that other users asked about similar images. The formulation of relevant questions is a complex problem because it may depend on factors external to the image itself, such as the user's intention or the context in which the image is analyzed. This type of knowledge is difficult to elicit and model but can be essential to provide a satisfactory user experience. For this reason, we hypothesize that a case-based reasoning approach will be useful for this problem because cases represent past experiences and might implicitly capture part of this knowledge. In this paper, we propose and evaluate the use of different similarities to retrieve similar images and propose a method to cluster and filter all the questions associated with those images so that we do not ask the same using different formulations.

The rest of the paper is organized as follows. Section 2 presents related work on VQA. Section 3 describes the proposed case-based approach to retrieve relevant questions about an image. Sections 4 and 5 present the dataset and the different measures of similarity between images that we used in the study. Section 6 describes the experiment that have been carried out and the results obtained. Finally, Section 7 discusses our conclusions and proposes future lines of work.

2 Related work

In the literature, VQA has been achieved using both knowledge-light [9, 20] and knowledge-intensive methods [18, 19]. Both approaches typically involve the following steps: 1) object detection in an image with high accuracy with fine-grained details; 2) language comprehension of the question; and 3) compilation of the answer utilizing the information from steps 1 and 2, and external knowledge sources (mostly in knowledge-intensive systems) [3]. Step 3 is the most challenging one due to the complex reasoning we have to carry out. Depending on the questions and the type of images, questions are very easy to answer from the objects detected in the image. However, often, there is a need to infer knowledge from other knowledge sources to find the coincidences between our question and the objects detected in our image [13]. This is especially relevant for open-ended or ambiguous questions, questions with multiple answers or questions that require additional information not available on the image [19]. Besides, knowledge-intensive approaches infer knowledge from other knowledge sources to disambiguate the uncertainties between our question and the objects detected in our image [13].

Some data-driven models encapsulate all three steps. Deep vision-language (VL) models that perform VQA using sequence-to-sequence models are trained on extremely large datasets. A key component of VL models is learning the alignment between images and paired text in an unsupervised manner using contrastive losses. Given an $< image, question >$ pair, VL models generate answers using auto-regressive sequence generation models by using the granular alignments between the question and the image as well as linguistic proficiency acquired in pre-training. Some recent models are the ALIGN model trained using contrastive loss [8], ALBEF model trained using the inter-modal alignment [9]; VL model trained using Triple Constrastive Loss [20]; and VinVL [21] which integrates improved object detection with VL modeling. Overall, these data-driven methods treat VQA as an end-to-end task instead of a pipeline of sub-tasks as in knowledge-intensive approaches. VQA is one of the many downstream tasks for a fine-tuned VL model. To support VQA, the VL model is further trained using a dataset of $< image, question, answer >$ triplets where the input is the $< image, question >$ pair, and the ground truth is the answer in textual format. VQA can be modeled as a generative task and a classification task. A generative VQA model will employ a text generator such as GPT [15] to generate the answer. Conversely, a classification model will employ a softmax layer at the end of the fusion encoder to predict the answer from a given set of answers [20].

While deep learning models are limited by the knowledge they have learned from the training data, knowledge-intensive methods [13, 19] get an answer using also external knowledge from different complementary knowledge sources apart from the images and the questions. Using additional knowledge sources can be remarkable in VQA tasks for *complex* or common sense questions. For example, in a question like "How many types of fruits are in the picture?". The algorithm should recognize not only the foods that we have in the image but also identify and understand which foods are fruits and which ones are not. Different approaches use complementary knowledge like the *Grounded Visual Question Answering* model (GVQA) [1] that use different algorithms depending on the type of question that we come across (yes/no or non-yes/no question). They also divide the model's behaviour process to get the answer in different steps (getting important parts in the image, retrieving concepts from the question, classifying the type of questions, or predicting the answer). VLC-BERT [16] is a VQA model that uses COMET (Commonsense Transformer) a commonsense reasoning generation transformer model that given a subject and a relation, predicts a possible object. An example from the authors is if the subject is "taking a nap" and the relation is "causes" a possible object is "have energy". It is trained and tested on ATOMIC [17] and ConceptNet knowledge graphs both of which consist of social commonsense knowledge. Commonsense reasoning extracted from the COMET is used in VLC-BERT to improve answer generation making it knowledge-intensive.

In this paper, we propose an approach using Case-based reasoning to reuse questions for similar images and avoid the use of other additional knowledge sources while maintaining a good performance of the question-answering process.

3 A Case-Based Reasoning model for VQA

The generation of relevant questions to describe an image is a complex problem because it may depend on factors external to the image itself, such as the user's intention or the context in which the image is analyzed. For example, a history student may visit an exhibition of classical painting to see historical figures and events, while another art student may visit the same exhibition paying special attention to the use of color and the organization of the elements in the scene. Each of these students would ask different questions about the same paintings because of their objective and, therefore, the type of information they seek is different.

This type of knowledge is difficult to elicit and model but can be essential to provide a satisfactory user experience. In this context, case-based reasoning techniques can be useful because cases, representing past experiences, capture, to some extent, this type of knowledge. We propose to describe images by retrieving and reusing questions that have been relevant to describe similar images in past situations.

The description of the *cases* is composed, in its simplest version, of an image. This description could be extended with knowledge about the user who interacted with the image and his context. The case solution consists, in its simplest version, of the questions that the user considered relevant to describe the image. This information could be extended with the answers to those questions or the user's reactions to the answers obtained.

When a user wants to get the description of an image, we search for similar cases in our experience base, to reuse the questions that were asked back then. Once the k most similar cases are retrieved, we must adapt the retrieved questions to the current context. This adaptation process must consider, in addition to the differences between the past situations and the current ones, the possible overlaps between questions retrieved from different similar cases.

In the current version of our system, cases contain images and their associated questions (see figure 1). The retrieval of the most similar cases is performed using different similarity measures between images, with lower or higher semantic load, which will be introduced in section 5. The adaptation process is in charge of selecting the most relevant questions from all the retrieved images. We must consider that being similar images, there will be many repeated questions (even if they are not formulated the same) that must be filtered. For example, from two similar images we could retrieve the questions "What time of the day is it?" and "Is it daytime or nighttime?", and we should not ask both. To solve this problem, we propose to generate embeddings of the questions using a language model, apply a clustering algorithm, and select the question closest to the centroid of each cluster. In our case, we use *gte-small* [10] to generate the embeddings, a language model based on BERT, and the DBSCAN [5] algorithm to generate the clusters.

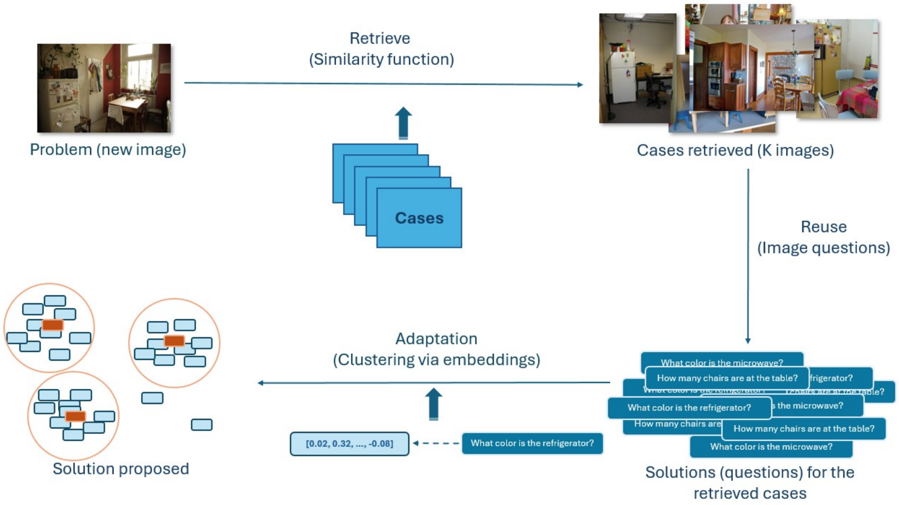

Retrieve
(Similarity function)
Problem (new image)
Cases
Cases retrieved (K images)
Reuse
(Image questions)
Adaptation
(Clustering via embeddings)
What color is the microwave?
How many chairs are at the table?
What color is the refrigerator?
How many chairs are at the table?
What color is the microwave?
[0.02, 0.32, ..., -0.08]
What color is the refrigerator?
Solutions (questions) for the retrieved cases
Solution proposed

Fig. 1. Question retrieval CBR system. The retrieval phase finds the K most similar images (based on a predefined metric) to the query image. Questions associated with these retrieved images are then extracted. The adaptation phase generates embeddings for these questions and groups them into clusters. Finally, the system proposes questions associated with the most representative embedding from each cluster.

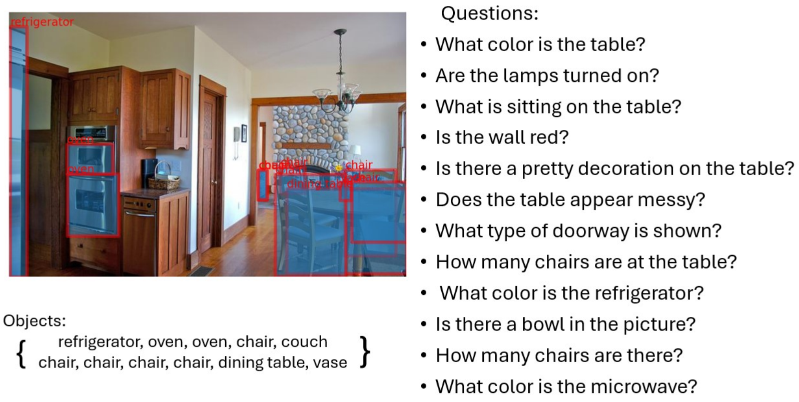
4 The Visual Question Answering dataset

The Visual Question Answering Dataset and Challenge[1] (VQA v2.0) [6] contains open-ended questions about images requiring an understanding of vision, language, and commonsense knowledge to answer. The dataset contains over 200k COCO dataset images, 5.4 questions per image on average and 10 correct answers for each question.

The images show a wide variety of scenes including landscapes, house interiors, everyday objects, wild and domestic animals, people doing different activities, etc. Each of the images contains information about the entities that appear in it, in particular their categories and their bounding boxes. There are 80 different types of entities (person, umbrella, dog, car, bed, clock...) [12]. Each image also has a variable number of questions associated with it (at least 3) and for each of them 10 answers in natural language. The questions, in turn, are tagged with a type from among 65 possible categories. The number of questions in each category is highly variable, the most numerous categories being *how many*, *is the*, *what*, *what color is the*, and *what is the*.

For example, figure 2 shows one of the images in the dataset and its associated information. In this case, it is a domestic kitchen in which 11 objects have been identified. The image has 12 associated short questions that will require different

[1] `https://visualqa.org/` (last accessed on 04/26/24)



Objects:

{ refrigerator, oven, oven, chair, couch
chair, chair, chair, chair, dining table, vase }

Questions:

- What color is the table?
- Are the lamps turned on?
- What is sitting on the table?
- Is the wall red?
- Is there a pretty decoration on the table?
- Does the table appear messy?
- What type of doorway is shown?
- How many chairs are at the table?
- What color is the refrigerator?
- Is there a bowl in the picture?
- How many chairs are there?
- What color is the microwave?

Fig. 2. Information associated with an image: objects that appear in the image (type and bounding box) and questions (and their categories). Each question has 10 answers that are not shown for clarity.

types of answers (yes/no, colors, numbers...). It is interesting to note that some of the questions cannot be answered with the labeled objects (*Is there a pretty decoration on the table?, Is the wall red?*).

In this work, we have selected a subset of the dataset consisting of 5000 images so that their questions follow a similar distribution to those in the original dataset in terms of question types as depicted in figure 3.

5 Similarity between images

Let us recall that our goal is to describe an image using questions that have been relevant to describe similar images in the case base. Therefore, the quality of the retrieved questions will depend on how we measure the similarity between images. Next, we introduce several similarity measures that attempt to capture different aspects that we can take into account when considering whether two images are similar. We will also show examples of similar and disimilar images to the target image in Figure 4 according to each of them.

5.1 Pixel similarity metric

Pixel-level image comparison is performed by comparing pixels at the same position in both images. This type of similarity measure can be useful, for example, when comparing landscape images where color plays a key role. Pixel similarities require that both images have the same size, so we resized the images in the dataset to 640 x 480 pixels, the modal size in the dataset, constituting 20.26% of the original images. There are different ways to compare two images based on

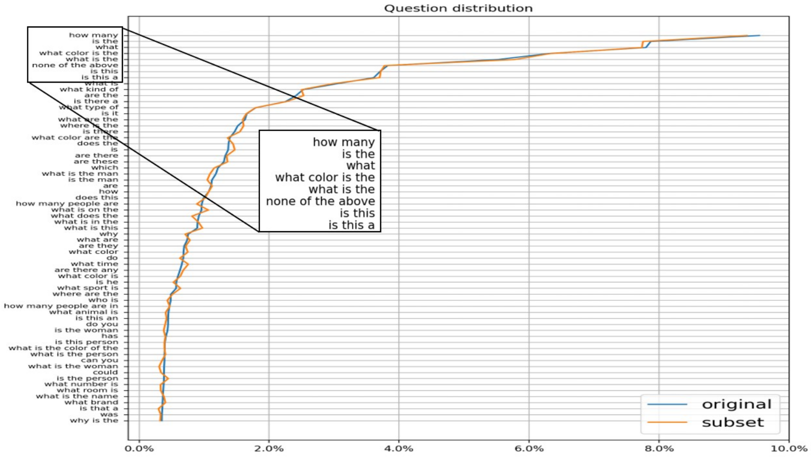


Fig. 3. Question type distributions.

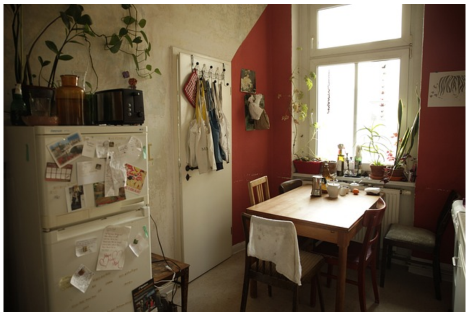


Fig. 4. Target image used to exemplify the different similarity measures.

the pixels composing them. We opted for mean square error between pixels due to its simplicity and interpretability.

Normalized Root Mean Squared Error This metric calculates the average squared distance between pixels in the same position, then takes the square root to get a value in the same color units. Then, values are normalized. The similarity is computed as 1 minus this distance.

$$sim(\mathbf{im_A}, \mathbf{im_B}) = 1 - \frac{\|\mathbf{im_A} - \mathbf{im_B}\|}{\|\mathbf{im_A}\|} \tag{1}$$

where $\mathbf{im_A}$, $\mathbf{im_B}$ are both images represented as pixel matrices and $||.||$ is the Frobenius norm. Figure 5 shows an example of similar and dissimilar images using this metric.

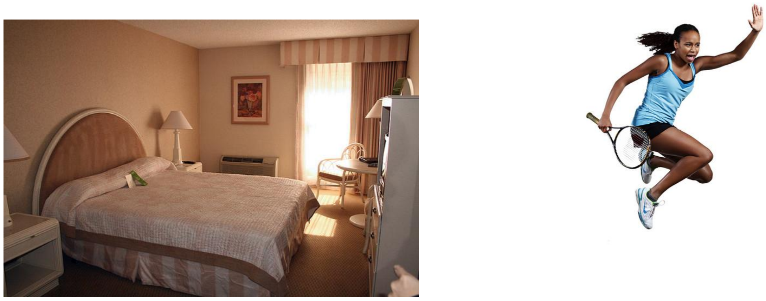


Fig. 5. Example of NRMSE based similarity calculation against image in figure 4: First image has a similarity of 0.4535 being the most similar in our dataset. The second image has a similarity of -0.4771 being the least similar in our dataset.

5.2 Object detection similarity metrics

We can consider two images to be similar when they show similar objects. The images in the dataset already contain annotations about the entities in them and their bounding boxes, but we need to use some object recognition model to use images external to the dataset. We delegate that task to the Retina Net [11] model, a convolutional neural network (CNN) trained on the COCO dataset. Next, we define two different similarity measures based on the entities appearing in the images, depending on whether we use their labels or their bounding boxes.

Tag-based similarity Given the set of object tags (without duplicates) in both images, we compute the similarity as the intersection over the union between the two sets.

$$sim(\mathbf{im_A}, \mathbf{im_B}) = \frac{|obj_A \cap obj_B|}{|obj_A \cup obj_B|} \tag{2}$$

where obj_A, obj_B are the sets of objects in both images and $|.|$ is the cardinality of the set. An example of similar and dissimilar images using this metric is shown in figure 6.

Bounding box-based similarity This similarity is intended to capture the idea that larger objects tend to have more importance in an image than smaller ones. First, we calculate how much of the image area the objects of each class occupy relative to the total area occupied by all objects. These values represent the relative importance of the object class in the image. Next, we sum the minimum value for each class in both images. It is easy to see that this value is between 0 and 1.

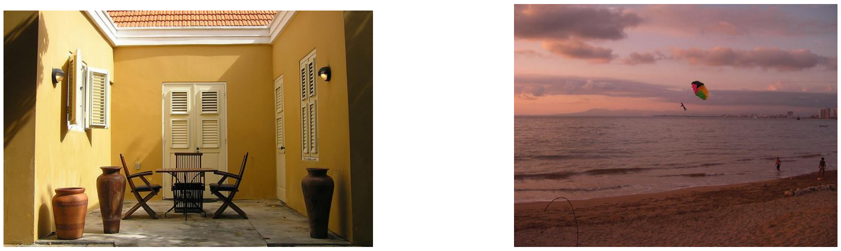


Fig. 6. Example of tag-based similarity calculation against the image in figure 4: The first image has a similarity of 0.75, being the most similar in our dataset. The second image has a similarity of 0, the lowest value.

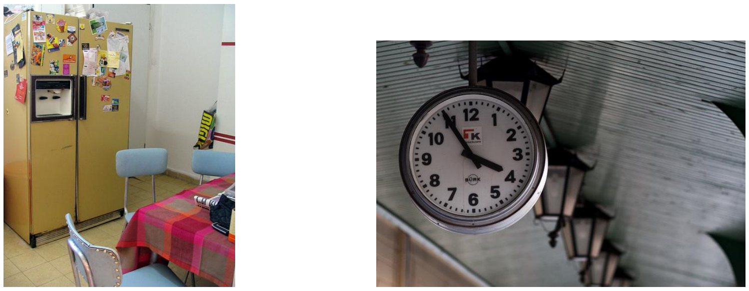


Fig. 7. Example of bounding boxes based similarity calculation against image in figure 4: First image has a similarity of 0.82898, being the most similar in our dataset. The second image has a similarity of 0, the lowest value.

$$sim(\mathbf{im_A}, \mathbf{im_B}) = \sum_{c \in C} min\left(\frac{r_A^c}{r_A}, \frac{r_B^c}{r_B}\right) \tag{3}$$

where C is the set of object classes defined in COCO, r_A^c (resp. r_B^c) is the sum of the areas of objects belonging to the class c in the image A (resp. B), and r_A is the sum of the areas of all the objects in the image A (resp. B). An example of similar and dissimilar images using this metric can be seen in figure 7.

5.3 Embedding-based similarity metric

Image embeddings are low dimensional representations in a *latent space* that capture, to some extent, semantic features of the images. They are extracted from internal layers of already trained neural networks and tend to group closer images with common features. Embedding-based similarities consider that two images are more similar the closer their embeddings are.

We use the Img2Vec model from the PyTorch library with the DenseNet [7] architecture to obtain a 1024 feature embedding of each image. DenseNet is a CNN architecture characterized by having dense connections between layers,

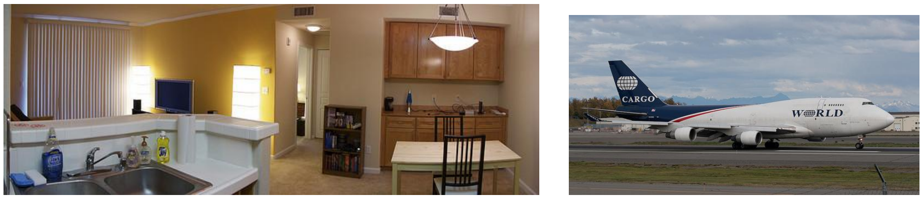


Fig. 8. Example of cosine similarity among image embeddings against the image in figure 4: The first image has a cosine similarity of 0.6958 being the most similar in our dataset. The second image has a cosine similarity of -0.0939 being the least similar in our dataset.

which means that each layer is directly connected to all subsequent layers in the model. This allows the flow of information through the different layers to be more fluent and direct, which can result in better performance and better generalization ability compared to other architectures. DenseNet architecture achieves state-of-the-art performance on various benchmark image classification datasets and it is also suggested it can be a good feature extractor for various computer vision tasks [7].

Cosine similarity Cosine similarity is particularly useful when working with high-dimensional data such as image embeddings because it considers both the magnitude and direction of each vector. This makes it more robust than other measures like Euclidean distance, which only considers the magnitude.

$$sim(\mathbf{im_A}, \mathbf{im_B}) = \frac{\mathbf{v_A} \cdot \mathbf{v_B}}{\|\mathbf{v_A}\| \|\mathbf{v_B}\|} \tag{4}$$

where $\mathbf{v_A}$, $\mathbf{v_B}$ are the embeddings of the images $\mathbf{im_A}$ and $\mathbf{im_B}$, $\cdot$ represents the dot product between them, and $||.||$ is defined as the norm of a vector. Figure 8 shows an example of similar and dissimilar images using this metric.

6 Experiment and results

The goal of the experiment is to assess the effectiveness of different similarity measures to retrieve images with semantically relevant associated questions. Our initial hypothesis are:

- Pixel similarity metrics, lacking semantic information, will exhibit poor performance in retrieving images with relevant questions. We expect to retrieve images with a high degree of pixel similarity but potentially low semantic relevance to the query.
- Object detection similarity metrics, while incorporating semantic information through object category recognition, will be limited by the predefined categories within the training dataset. Additionally, we believe that images with few or no objects to detect, such as landscapes, will demonstrate a weakness of these metrics.

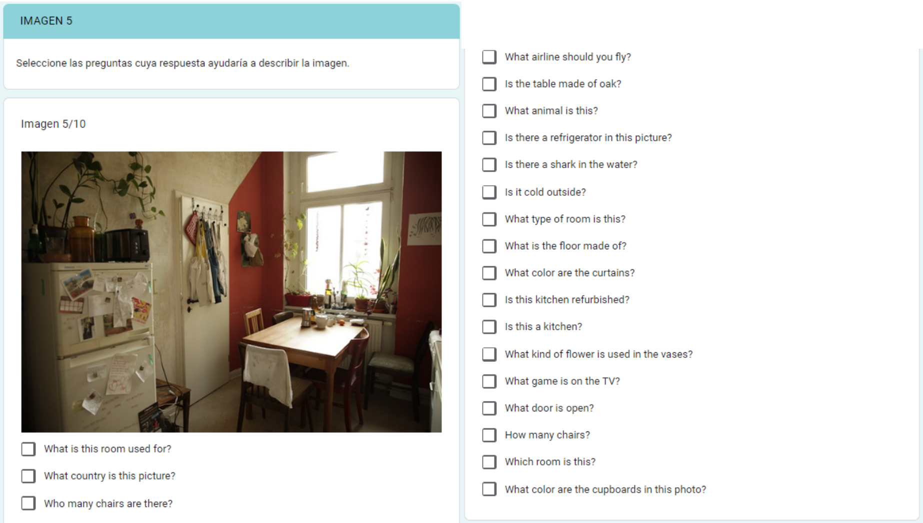


IMAGEN 5

Seleccione las preguntas cuya respuesta ayudaría a describir la imagen.

Imagen 5/10

□ What is this room used for?

□ What country is this picture?

□ Who many chairs are there?

□ What airline should you fly?

□ Is the table made of oak?

□ What animal is this?

□ Is there a refrigerator in this picture?

□ Is there a shark in the water?

□ Is it cold outside?

□ What type of room is this?

□ What is the floor made of?

□ What color are the curtains?

□ Is this kitchen refurbished?

□ Is this a kitchen?

□ What kind of flower is used in the vases?

□ What game is on the TV?

□ What door is open?

□ How many chairs?

□ Which room is this?

□ What color are the cupboards in this photo?

Fig. 9. Example of image and questions in the conducted experiment.

– Embedding-based similarity metrics, in the same way as object detection, incorporate semantic information. We hypothesize they will perform more consistently across a wider range of image types than pixel-based or object-detection-based metrics.

6.1 Experiment set-up

We selected 20 new images from Pixabay[2] with different topics (animals, landscape, street, transport, technology, rooms, sports, and food) so that they would not be related to the ones present in the case base. Then we used the 4 similarity metrics introduced in the previous section (pixel-based, tag-based, bounding box-based, and embedding-based) to retrieve 4 questions associated with each one of them. The retrieval was performed using the 50 most similar images according to each similarity and the clustering method described in Section 3. We also introduced a baseline composed of the 4 most typical types of questions in the entire dataset (we used clustering over all the questions and selected representatives of the 4 largest clusters). Therefore, the experiment was performed with 20 images and a pool of 20 randomly ordered questions for each of them: 4 questions using each similarity and 4 baseline questions. An example is shown in figure 9.

The 20 images were divided into 2 online tests with 10 images each to enhance participation so that each test could be completed in less than 10 minutes. The participants were asked to choose among the pool of questions those *whose answers they considered would help to build a description of the given image.*

[2] `https://pixabay.com/` (last accessed on 04/26/24)

Similarity	**Questions selected(%)**
Baseline	6.43%
Pixel-based (NRMSE)	17.78%
Tag-based (RetinaNet)	48.22%
Bounding box-based (RetinaNet)	54.58%
Embedding cosine (DenseNet)	65.53%

Table 1. The percentages shown represent the selectivity of each model, rather than an overall success rate. These values indicate the proportion of retrieved questions that users chose for each model. A score of 100% would signify that all users selected all questions presented by that particular model.

The participants could select as many questions as they wished, and they knew nothing about where those questions came from.

6.2 Experiment results

A total of 80 people answered to test A and 68 to test B. The average number of questions selected per person and image was 7.08 out of 20. The evaluation was carried out by analyzing the selection rates of questions retrieved using each similarity metric. This way, we aimed to identify the model that consistently retrieved the most relevant questions. The results are summarized in Table 1.

As expected, the naive approach of selecting the most common questions in the dataset does not work for this type of problem. Pixel-based similarity is a little better but not much more than the baseline. The similarity metrics that consider the semantic information contained in the image are a better choice. In the detected object similarities is interesting that the size of the objects in the picture seems to be relevant. Finally, the similarity based on embeddings is the clear winner. We believe that leveraging DenseNet embeddings for image retrieval goes beyond simple object recognition and likely captures additional features that contribute to semantic similarity (as stated in section 5.3), potentially leading to the retrieval of more relevant images.

It is interesting to analyze the results by dividing the set of test images into different categories. A first distinction can be made between indoor and outdoor images. Within the indoor images we can distinguish between those inside rooms and those focusing on specific items. As for outdoor images, they can be divided into images of nature and images of the urban environment. We analyze the results of the experiment according to this classification as follows (see Table 2):

- **Indoor room images (4 test images)**. Similarities based on object detection perform reasonably well because the images contain several recognizable objects such as chairs, beds, or animals. On the other hand, pixel-based similarities are usually not useful because of the diversity of sizes and, especially, colors of these objects. Similarity based on embedding is the most effective in most cases, except for one of the images in this category.

	Questions selected (%)					
	Interior			**Exterior**		
Similarity	**Room**	**Items**	**All**	**Urban**	**Nature**	**All**
Baseline	16.54%	5.37%	10.34%	2.48%	5.27%	3.24%
Pixel-based (NRMSE)	17.28%	4.49%	10.18%	21.32%	31.13%	24.00%
Tag-based (RetinaNet)	51.10%	46.99%	48.82%	54.04%	30.88%	47.73%
Bounding box-based (RetinaNet)	66.54%	53.09%	59.07%	54.46%	41.42%	50.90%
Embedding cosine (DenseNet)	76.38%	61.62%	68.18%	62.50%	65.68%	63.37%

Table 2. Selection rate for each model divided by test image type. Analogous to Table 1.

- **Indoor images focused on specific items (5 test images)**. Similarity based on embedding again performs best in this group. Similarities based on object detection are limited by the predefined categories and objects that the model can recognize, and that is especially important for these images. Pixel-based similarity performs poorly in this group, probably due to the variability of colors in everyday objects.
- **Images of urban environments (8 test images)**. In this group, we appreciate two types of images. Object detection-based similarity obtains the best results only in some images that are not saturated with objects and certain elements are clear protagonists. Otherwise, embedding-based similarity performs better probably because it considers more objects that are important in the image. Pixel-based similarity only works well when the tone of the image is important for meaning (rainy days will have more grayish tones, and sunny days will have warmer tones).
- **Images of nature environments (3 test images)**. Similarities based on object detection perform worse in this group than in the previous ones because there are usually not many objects in open, outdoor landscapes, although they improve if animals are present. Pixel-based similarity performs better in images where the color is very uniform, such as the sea, the sky or a meadow, although it still performs worse than bounding-box-based and embedding-based similarities. Embedding-based similarities continue to be the top performers with this type of image.

Our evaluation demonstrates that embedding cosine similarity offers the best sim metric for question recommendation. This approach retrieved a user-reported suitability rate exceeding 60%. Object-based similarity metrics achieved a modest performance when the model identified recognizable objects within the images. However, this approach exhibited limitations with landscape imagery or scenes lacking identifiable object categories. Notably, incorporating relative object size improved object-based similarity by over 10% in certain scenarios. Pixel-based similarity metrics consistently underperformed, even operating below baseline results in some cases.

7 Conclusions and Future Work

In this paper we have proposed a case-based reasoning approach to help users, especially those with visual impairments, to formulate relevant questions about an image based on questions that other users previously asked about similar images. In particular, we evaluated the use of different similarity measures to retrieve similar images and proposed a method to cluster and filter the questions associated with those images so that the same question is not asked multiple times using different formulations.

The results of our user experiment suggest that pixel-based similarities do not suit this type of problem, except for images where color or hue is a dominant semantic feature (e.g., snowy landscapes or water scenes). Similarity based on detected objects works well when the image contains a few important elements that can be recognized by the detection model but fails when there are too many elements in the image, none, or the detection model is not able to recognize the relevant entities. Embedding-based similarity depends on the quality of the embeddings (and thus the model and dataset used to create them) but seems to work better on a wider range of image types. Embeddings are a very effective way to encode the general semantic information present in the image, but they are also less transparent and difficult to interpret. For example, the embedding-based similarity performed worse than expected in one image of the experiment, but we cannot explain why this is the case. These results encourage us to study the possible combination of different similarity measures to improve the relevance of the retrieved images in different situations. This is an interesting and challenging line of research, as it is not straightforward to characterize the information used when calculating similarity based on embeddings.

There are several lines of research to improve this work. One of the most important problems that we have not yet addressed concerns how to adapt the retrieved questions to the current image. Two different images could show similar scenarios with different contexts and that could make some questions relevant to the first image absurd for the second one. We also used a very simple representation of the cases in this work that could be enriched with knowledge about the user who asked the question, such as her intention and context (physical, temporal, social, etc.), and her reactions to the answers obtained. All this new information could be taken into account to retrieve more relevant cases and make more complex adaptations. Finally, once the relevant questions about the image have been retrieved, we should compose a description based on the answers to those questions and evaluate whether that description is useful for visually impaired people.

Acknowledgements This work has been partially funded by the Ministry of Science, Innovation and Universities (PID2020-114596RB-C21, PID2021-123368OB-I00) and the Complutense University of Madrid (Group 921330).

References

1. Agrawal, A., Batra, D., Parikh, D., Kembhavi, A.: Don't just assume; look and answer: Overcoming priors for visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4971–4980 (2018)
2. Antol, S., et al.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
3. Basu, K., Shakerin, F., Gupta, G.: Aqua: Asp-based visual question answering. In: Practical Aspects of Declarative Languages: 22nd International Symposium, PADL 2020, New Orleans, LA, USA, January 20–21, 2020, Proceedings 22. pp. 57–72. Springer (2020)
4. Caro-Martínez, M., Wijekoon, A., Díaz-Agudo, B., Recio-García, J.A.: The current and future role of visual question answering in explainable artificial intelligence. In: Malburg, L., Verma, D. (eds.) Proceedings of the Workshops at the 31st International Conference on Case-Based Reasoning (ICCBR-WS 2023) co-located with the 31st International Conference on Case-Based Reasoning (ICCBR 2023), Aberdeen, Scotland, UK, July 17, 2023. CEUR Workshop Proceedings, vol. 3438, pp. 172–183. CEUR-WS.org (2023), `https://ceur-ws.org/Vol-3438/paper_13.pdf`
5. Ester, M., Kriegel, H., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Simoudis, E., Han, J., Fayyad, U.M. (eds.) Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA. pp. 226–231. AAAI Press (1996), `http://www.aaai.org/Library/KDD/1996/kdd96-037.php`
6. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6904–6913 (2017)
7. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. pp. 4700–4708 (2016)
8. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning. pp. 4904–4916. PMLR (2021)
9. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems **34**, 9694–9705 (2021)
10. Li, Z., Zhang, X., Zhang, Y., Long, D., Xie, P., Zhang, M.: Towards general text embeddings with multi-stage contrastive learning. CoRR **abs/2308.03281** (2023). https://doi.org/10.48550/ARXIV.2308.03281, `https://doi.org/10.48550/arXiv.2308.03281`
11. Lin, T., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. CoRR **abs/1708.02002** (2017), `http://arxiv.org/abs/1708.02002`
12. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P.: Microsoft coco: Common objects in context (2015)
13. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition. pp. 3195–3204 (2019)

14. Patil, A.P., Behera, A., Anusha, P., Seth, M., Prabhuling: Speech enabled visual question answering using LSTM and CNN with real time image capturing for assisting the visually impaired. In: TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), Kochi, India, October 17-20, 2019. pp. 2475–2480. IEEE (2019). https://doi.org/10.1109/TENCON.2019.8929263, `https://doi.org/10.1109/TENCON.2019.8929263`
15. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
16. Ravi, S., Chinchure, A., Sigal, L., Liao, R., Shwartz, V.: Vlc-bert: Visual question answering with contextualized commonsense knowledge. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1155–1165 (2023)
17. Sap, M., Le Bras, R., Allaway, E., Bhagavatula, C., Lourie, N., Rashkin, H., Roof, B., Smith, N.A., Choi, Y.: Atomic: An atlas of machine commonsense for if-then reasoning. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 3027–3035 (2019)
18. Wang, P., Wu, Q., Shen, C., Dick, A., Van Den Henge, A.: Explicit knowledge-based reasoning for visual question answering. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence. pp. 1290–1296 (2017)
19. Wu, Q., et al.: Visual question answering: A survey of methods and datasets. Computer Vision and Image Understanding **163**, 21–40 (2017)
20. Yang, J., Duan, J., Tran, S., Xu, Y., Chanda, S., Chen, L., Zeng, B., Chilimbi, T., Huang, J.: Vision-language pre-training with triple contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15671–15680 (2022)
21. Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., Gao, J.: Vinvl: Revisiting visual representations in vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5579–5588 (2021)