IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Smart Visual and Question And Answering System Using ArtificialIntelligence

1st Anshul Kakroo 2nd Shivam Magar 3rd Dipak Patil 4th Yash Rajoriya 5th Pradnya Kasture.

Dept. of Computer Science (RMDSSOE) Savitribai Phule Pune University Pune, India

Dept. of Computer Science (RMDSSOE) Savitribai Phule Pune University Pune, India

Dept. of Computer Science (RMDSSOE) Savitribai Phule Pune University Pune, India

Dept. of Computer Science (RMDSSOE) Savitribai Phule Pune University Pune, India

Dept. of Computer Science (RMDSSOE) Savitribai Phule Pune University Pune, India

(Student, Department of Computer Engineering, Rmdssoe, Pune, Maharashtra, India)

Abstract: **The development of image-related intelligent interaction technology has been accelerated by the use of artificial intelligence methods in the field of image processing. By posing questions that are relevant to the image, visual question answering (VQA) gathers data that will help people better understand the images. A great deal of study has been done in each of their separate domains. Vision and language are the two fundamental components of human intelligence to grasp the real world and also the basic components to realise artificial intelligence. The development of visual question-answering technology across the visual field and natural language disciplines has recently been a research hotspot due to the ongoing promotion and use of deep learning in the fields of computer vision and natural language processing. By asking questions that are pertinent to the image's content, visual question answering (VQA) for intelligent engagement gathers image data with the ultimate goal of enhancing image comprehension. The visual question answering system faces enormous hurdles as a new area of research, thus we must study and go deeper into this area.**

INTRODUCTION

VQA is a new and extremely challenging area of research in the field of AI. It encompasses disciplines linked to computer vision. In order for the model to automatically predict responses to questions involving photos and images-related queries, it must develop a model to analyse images and understand the problem. Since the VQA challenge in 2014, a wide range of visual question and answer formats have been proposed.

Depending on whether external knowledge bases are used, the present VQA models are divided into two groups: knowledge base-based models and joint embedding models.

Basic models of visual question answering frequently use linear classifiers or multi-layer perceptual (MLP) classifiers to link image vectors with text properties [9–11]. It is utilised the Turing Test (Turing Test). was first put forth by Alan Turing in 1950 and is used to test whether a machine can show intelligence equivalent to or indistinguishable from humans. To find out whether a machine is intelligent, perform the Turing test. Many question-answering systems have been created as a result of computers being used to answer inquiries from people, furthering the field of natural language processing technology. In 2015, A. Agrawal created

a visual question-answering challenge that provides a visual question-answering system with a picture and a natural language query. It is your responsibility to react in clear, everyday language.

Malinowski et al. [12] were the first to develop a hybrid embedding model dubbed Neural-Image-QA that was applied to real-world scenes.

Neural-image-QA uses convolutional neural networks to extract picture information. Long short-term memory (LSTM) then creates the word order for the answer using the extracted feature vector and the query text. The model's accuracy on the DAQUAR [13] dataset is 19.43%. The CNN+RNN foundational paradigm was often applied by later researchers as well. Zhou et al. while addressing the query text.

The bag-of-words model BOW, which is less complex than the long short-term memory LSTM and performs well with the VQA dataset, was offered as an alternative to the ffiOWIMG model W and transferred from the pre-trained Google Net to extract image features. Gao et al. used two different LSTM networks because they believed the query and the answer had different grammatical structures to encode the inquiry and decode the response and they follow.

The convolutional neural network was subsequently combined to produce the MQA model. M. Lin et al. produced joint feature vectors using a multimodal convolutional layer, presented a dual CNN model, and employed a convolutional neural network CNN to both encode the content of the images and extract the features of the question language.

Related works

Addressing questions visually recently, numerous datasets and techniques have been presented [15, 4, 33, 24, 2, 25, 14, 29], ranging from confined environments [15, 24, 29] to freeform natural language questions and replies [4, 33, 2, 25, 14]. For instance, [15] suggests employing a set vocabulary of objects, properties, and interactions between objects to create binary questions from templates. In order to react to questions about videos, [33] has researched cooperative parsing of videos and associated text. [24] investigated VQA using artificial (templated) and human-generated questions, with the restriction that the answers could only be to 894 different categories or sets of 894 different categories of objects and 16 different colours. Numerous publications published recently [2, 14, 25, 29] suggested neuronal.

LSTMs and CNNs are the two types of network models used in VQA. [2] introduced numerous natural VQA models as well as a sizable dataset for free-form and open-ended VQA.

[4] employs workers from the crowd to respond to queries from users who are visually challenged about visual information.

Adding data. Traditional data augmentation methods (such mirroring and cropping) have been utilised frequently recently [18, 31] to provide high capacity models more material to draw upon. The label distribution in the training data should remain mostly unchanged as a result of these changes. Only 6% of test questions are unsuitable for this modification. In order to ensure that every question in our dataset receives the same number of "yes" and "no" responses, we use human participants to gather extra scenarios. Our method can be thought of as semantic data augmentation in that sense. Many classification datasets, including ImageNet [7], make an effort to be balanced. Due to the substantial amount of notions that language may represent, this is however impossible for the VQA assignment on real photographs. This inspires us to employ abstract scenes.

Methodology

Language plus abstract visuals. Many studies have used abstract scenes to concentrate on high-level semantics and explore the relationship between it and other modalities like language [23, 16, 38, 40, 39, 3, 13, 34], including ones that automatically describe abstract scenes [16], create abstract scenes that depict a description [39], capture common sense [13, 23, 34], learn models of fine-grained interpersonal interactions [3], and discover the semantic significance of visual features [38, 40]. Some of these works have also used visual abstraction to "control" the distribution of data, as shown in [3]'s collection of an equal number of verb/preposition examples, and [38]'s use of different scenes to represent the same sentence or caption. In a similar manner, we balance the dataset by ensuring that each inquiry in the To the extent possible, the dataset has a scene for "yes" and a different scene for "no".

visual assurance. Men, rides, and elephants are examples of common-sense claims [30, 34] that are supported by real imagery [30] and abstract sceneries [34] respectively. We, on the other hand, concentrate on visually supported image-specific queries, such as "Is the man in the picture riding an elephant?" [39] maps the relationships between two items' visual attributes and further justifies these relationships.

They take a description as input and automatically create a scenario that is both believable and compatible with all of the tuples in the description. In this instance, we are interested in determining whether a single tuple (the question's summary) is present in a particular image or not, with the intention of responding to a general "yes/no" query regarding the picture.

Visual attention entails looking for and focusing on pertinent image areas. For the creation of image captions, [20, 36] employs alignment/attention. The only input is an image, and they attempt to use phrases and sentences to describe the entire image as well as specific local areas. We focus on a different issue: answering questions visually. An image and words (a question) are provided as input. In order to extract specific visual details from the areas of the image being discussed in the text, we want to align portions of the question to regions in the image.

VQA Datasets used in:

Real images: The MS COCO dataset was gathered to find images containing multiple objects and rich contextual information. Given the visual complexity of these images, they are well-suited for our VQA task. The more diverse our collection of images, the more diverse, comprehensive, and interesting the resultant set of questions and their answers.

Questions. Collecting interesting, diverse, and well-posed questions is a significant challenge. Many simple questions 1. https://github.com/tylin/coco-ui 4 may only require low-level computer vision knowledge, such as "How many cars are present in the scene?". However, we also want questions that require common sense knowledge about the scene, such as "Is there any objects a in the image?". Importantly, questions should also require the image to correctly answer and not be answerable using just common-sense information.

Working: This uses a classical CNN-LSTM model like shown below, where Image features and language features are computed separately and combined together and a multi-layer perceptron is trained on the combined feature

Working

A Visual Question Answering (VQA) system that incorporates Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks can effectively process both image and textual information to answer questions about images.

Here is a general outline of how a CNN-LSTM-based VQA system works:

1. Image Processing: The input image is passed through a CNN, such as the popular VGGNet or ResNet, to extract high-level visual features. The CNN transforms the image into a fixed-length feature vector, capturing the visual content and spatial information.

2. Question Processing: The input question is encoded into a fixed-length vector representation using an LSTM network. The LSTM sequentially processes the words in the question, capturing the semantic meaning and contextual information.

3. Fusion: The visual features from the CNN and the textual features from the LSTM are combined using a fusion mechanism. Various fusion approaches can be employed, such as element-wise addition, concatenation, or multiplicative interactions. The goal is to merge the visual and textual information into a joint representation.

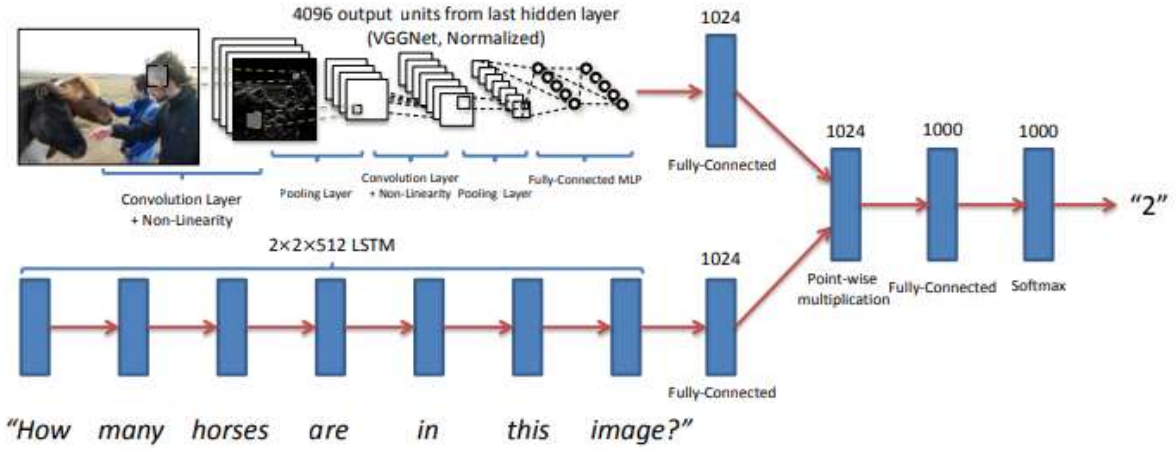
4. Answer Generation: The fused representation is then fed into a classifier or a fully connected network to predict the answer to the given question. The classifier may use softmax activation to output a probability distribution over a predefined set of answer choices. The answer with the highest probability is selected as the final answer.

5. Training: During training, a large dataset of image-question-answer triplets is used to learn the parameters of the CNN, LSTM, fusion mechanism, and the answer classifier. This is typically done using techniques such as backpropagation and gradient descent to minimize the loss between the predicted answers and the ground truth answers.

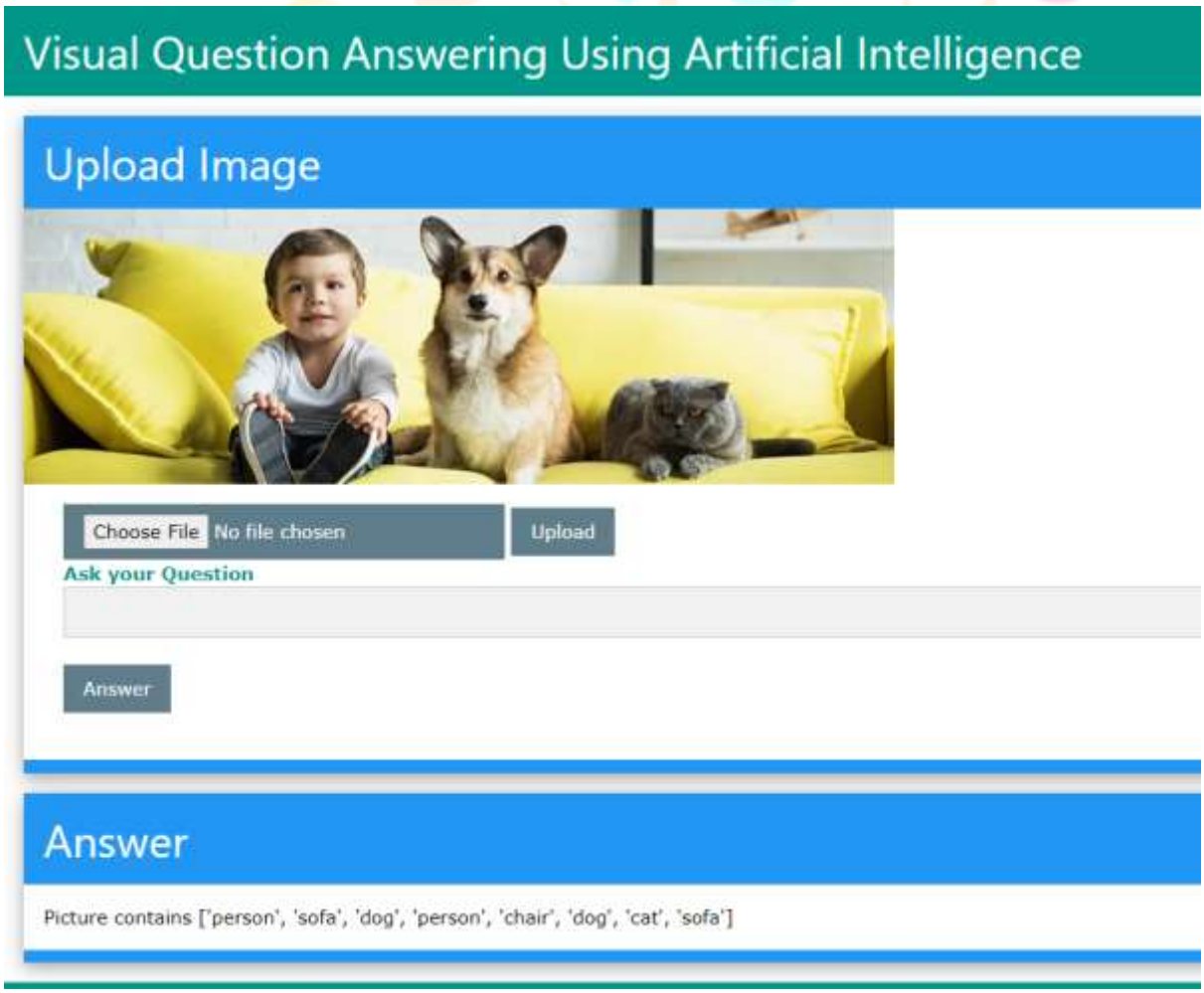
6. Inference: During inference or testing, the trained model is used to process new images and questions. The image is fed into the CNN for feature extraction, and the question is encoded using the LSTM. The fusion and answer generation steps are then performed to obtain the predicted answer.

By combining the visual processing capabilities of CNNs with the sequential reasoning abilities of LSTM networks, the VQA system can effectively understand the content of an image and generate meaningful answers to questions about that image.

It's worth mentioning that there are various techniques and architectures that can be explored to improve the performance of VQA systems, including attention mechanisms, multi-modal fusion methods, and more advanced network architectures.



Question is "what are objects present in tis picture?"



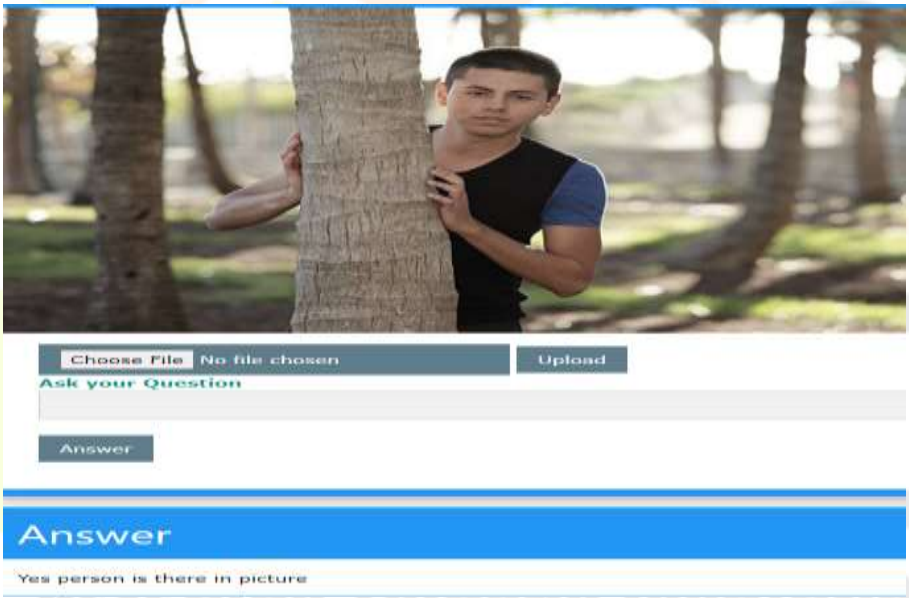
Experimental examples:

Results

We discover that despite the visual similarity of pairs of scenes with opposing ground truth answers to the identical questions, our model correctly predicts the responses for both scenarios. Additionally, we observe that our model has mastered the skill of focusing on the parts of the scene that appear to match to the parts that are most crucial to resolving the current query. Demonstrating visual knowledge is being able to (accurately) predict various outcomes to scenes that are slight semantic perturbations of one another**.**

Question is "is there monkey in this picture?"

Question is "is there any person in this picture?"



Conclusion

We make progress in this paper in automating the process of visual question answering. We focus on the issue of responding to binary inquiries about visuals. By adding complimentary scenes to the existing real images VQA dataset, we balance it so that almost all of the questions have a "yes" or "no" response for one scene and one of its closely related scenes. An technique needs to comprehend the image in order to perform effectively on this balanced dataset. We'll release our balanced dataset to the general public.

We provide a method that determines the area in the scene it should focus on, extracts a succinct description of the question in tuple form, and confirms the existence of the visual notion given in the question tuple in order to provide a response. On the balanced dataset, our technique significantly beats the language prior baseline and a cutting-edge VQA approach. We also provide qualitative findings that demonstrate how our method pays attention to pertinent areas of the scene in order to provide an answer.

Discussion

The development and implementation of a visual question answering (VQA) system have led to significant advancements in computer vision and natural language processing. A VQA system aims to answer questions about an image, bridging the gap between visual content and textual information. The discussion regarding the results of a VQA system would depend on various factors such as its accuracy, performance, and usability. Here are some possible outcomes and observations that could arise from such a discussion:

1. Accuracy: The accuracy of a VQA system is about 56.25% because some times it can not predict the correct output and sometimes model shows the errors in the system even if all the images contains object etc.

2. Performance: Performance of the model to the speed and efficiency of the VQA system in processing images and generating answers is faster and it can be used in real time applications. the system need some improvements in the dasets in model.

3. Interpretability: A VQA system can explain how it arrived at a particular answer is more trustworthy and useful.

Future Directions:

1. Human-Computer Interaction: By enabling more intuitive and natural communication, VQA systems can improve human-computer interaction. Users don't have to limit themselves to text-based searches when asking questions concerning visual content. This can be especially helpful in fields like augmented or virtual reality, smart home automation, and virtual or virtual assistants.

2. Assistive Technology: VQA systems can help people with disabilities or visual impairments by giving them access to visual data. These systems can help people who are blind or visually impaired learn and comprehend information by providing answers to questions based on visuals. This has uses in daily activities, navigation, and education.

3. Content-Based Image Retrieval: To enable more precise and accurate searches, VQA systems can be combined with image retrieval systems. Users can use natural language to describe an image they're looking for, and the system will return images that match the query. This has advantages in e-commerce, image databases, and visual search engines, among other things.

4. Autonomous Vehicles: VQA systems have the potential to significantly improve the perception and judgement abilities of autonomous vehicles. These systems can enhance the awareness of the environment, making autonomous cars safer and more effective. They do this by responding to questions about the environment or instantly detecting things.

5. Healthcare: VQA systems can be applied in medical imaging analysis, assisting healthcare professionals in interpreting medical images such as X-rays, MRIs, or histopathological slides. These technologies can help with quicker and more precise diagnosis by providing information on anomalies, structures, or diagnoses.

6. Education: VQA systems can be used to support learning and assessment in educational contexts. Questions from students on visual material allow for individualised and interactive learning. VQA can also be used in automated grading systems to provide immediate feedback on tasks or exams that rely on visuals.

7. Social Media and information Moderation: VQA systems can help identify potentially dangerous or improper information on social media platforms by analysing and responding to queries regarding photos or videos. This may assist in filtering out spam, hate speech, or explicit material.

8. Robotics and Industrial Automation: To improve perception and object identification skills, VQA systems can be linked into robotics and industrial automation systems. Robots can make intelligent decisions and carry out difficult activities in a variety of settings, such as manufacturing, logistics, or home support, by providing information about items or environments.

Appendix

Appendix1 : "What is/are" Analysis

For both actual photos and abstract scenes, we display the distribution of questions beginning with "What are" by their first five words. Take note of the variety of objects mentioned in the questions as well as the relationships between objects, such as "holding" and "sitting on". The range of responses to "What is" queries with various word endings is depicted . For instance, answers to queries that conclude in "eating" include "pizza," "watermelon," and "hot dog."

For example, inquiries that finish in "for?" or "picture?" have a wide range of responses. In other cases, the answer to a query like "holding?" is intuitive, "umbrella**."**

Appendix2 :"AGE" and "COMMONSENSE" of our model

Using VQA testdev accuracies, we choose our best model (deeper LSTM Q + norm I), and estimate its age and commonsense level. On the subset of questions in the VQA validation set for which we have age annotations (how old a human needs to be to answer the question correctly), we compute a weighted average of the average age per question, weighted by the accuracy of the model's predicted answer for that question. For the subset of questions in the VQA validation set for which we have commonsense annotations (whether positive or negative), we compute a weighted average of the average degree of commonsense per question, weighted by the accuracy of the model's predicted answer for that question.

Appendix3 : Issue of negation

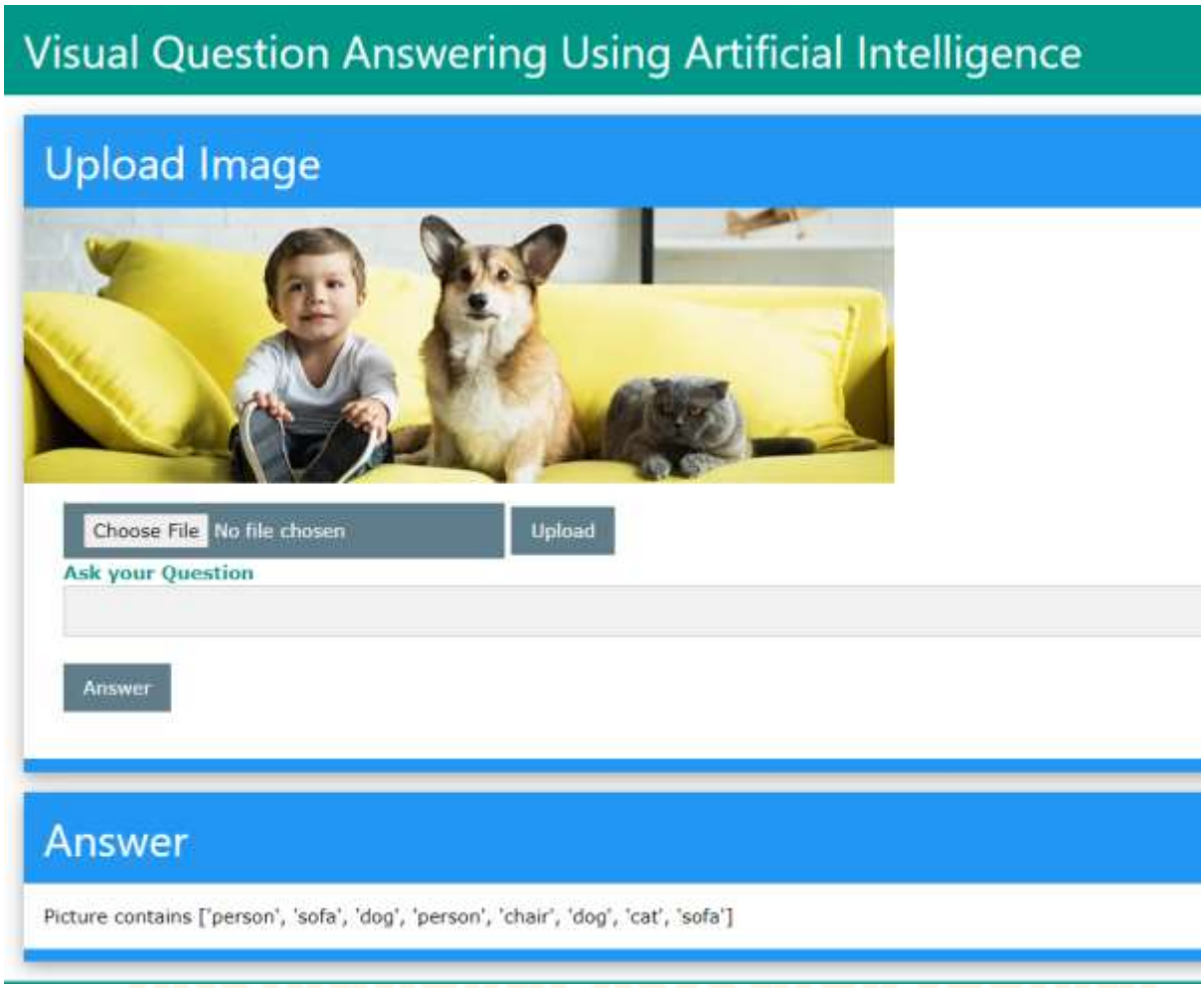
Our method causes a problem with how poorly negative questions are handled because it concentrates on the question's significant words. The extracted tuple, for instance, would be the same for the questions "Is the cat on the ground?" and "Is the cat not on the ground?," although their replies should ideally be the opposite. Because such negative statements are so uncommon in human speech, this topic of negation is difficult to address even in NLP research. For instance, less than 0.1% of the binary questions in the VQA training dataset contain the words "not," "isn't," "aren't," "doesn't," "don't", "didn't", "wasn't," "were," "shouldn't," "couldn't," or "wouldn't."

Appendix4 : Qualitative results

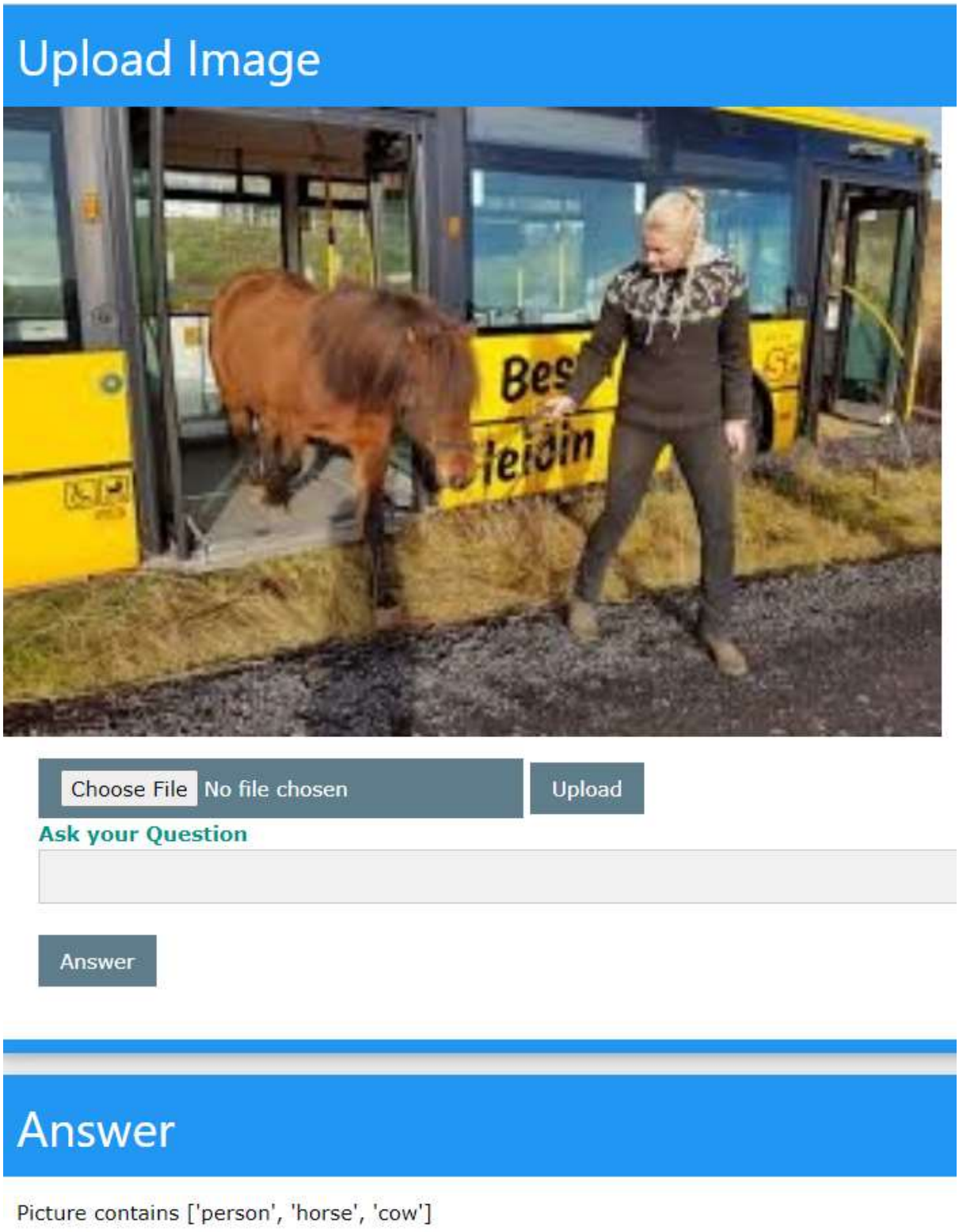
We show some qualitative results of our approach in including some failure cases. In each scene, the primary and secondary objects have been marked in red and blue boxes respectively.

Some results of model :

Question is "what are the objects present in this picture?"



Question is "What are the objects present in this image?"



References

[1] K. Simonyan and A. ZISSERMAn, "Very Deep Convolutional Networks for Large-Scale Image Recognition Computer Vision and Pattern Recognition," p. 1556, 2014, https://arxiv. org/abs/1409.1556.

[2] K. He, X. Zhang, S. Ren, and S. Jian, "Deep Residual Learning for Image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778, New york, NY, USA, December 2016.

[3] J. Donahue, L. Anne Hendricks, S. GUADARRAMA, V. Subhashini, and R. Marcus, "Long term recurrent convolutional networks forvisual recognition and description," in Proceedings of the IEEE Conference on Computer Vision and PatternRecognition, pp. 2625–2634, September 2015.

[4] K. Simonyan and A. Zisserman, "Two Stream Convolutional Networks for Action Recognition in videos," Computer Vision and Pattern Recognition, 2014, https://arxiv.org/abs/1406. 2199.

[5] J. Redmon, S. Divvala, and R. GIRSHICK, "You Only Look once: Unified,real-Time Object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788, Las Vegas, NV, USA, December 2016.

[6] S. Ren, K. He, R. B. Girshick, and C. Faster Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137–1149, 2017.

[7] A. Singh, V. Natarajan, M. Shah et al., "Towards VQA Models that Can Read," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Piscataway, pp. 8317–8326, IEEE Press, Long Beach, CA, USA, January2019.

[8] A. F. Biten, R. Tito, A. Mafla et al., "Scene text visual question answering," in Proceedings of the IEEE/CVF International Conference on Computer Vi-Emerging Telecommunications Technologies, vol. 31, no. 5, Article ID e3734, Seoul, Republic of Korea, Febrary 2020.

[9] B. Zhou, Y. Tian, S. Sukhbaatar, S. Arthur, and R. Fergus, "Simple Baseline for Visual Question Answering. Computer Vision and Pattern Recognition," 2015, https://arxiv.org/abs/ 1512.02167.

[10] K. Kafle and C. Kanan, "Answer-type Prediction for Visual," in Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, December 2016.

[11] C. Szegedy, V. Vanhoucke, and S. Ioffe, "question answering," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4976–4984, San Juan, PR, USA, 2016.

[12] S. Antol, A. Agrawal, and J. Lu, "Visual question answering," in Proceedings of the IEEE International Conference on Computer Vision, pp. 2425–2433, 2015.