International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 08 | August -2020 ISSN: 2582-3930

Visual Question Answering System (ViQAS)

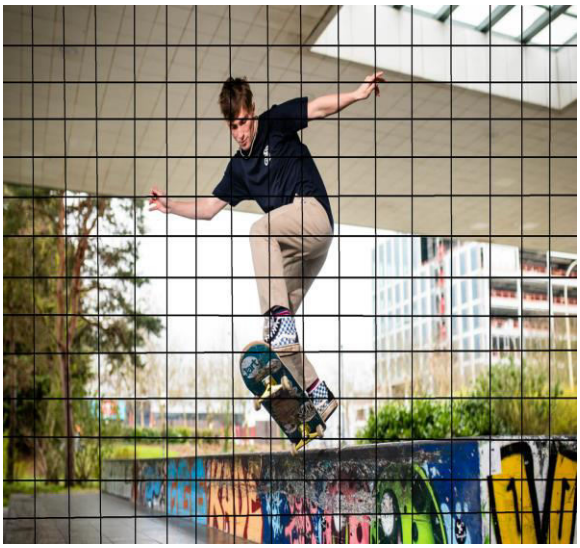
Abhishek Dutta Rishav Bhardwaj Shubhank Mehar Yasher Mahmood

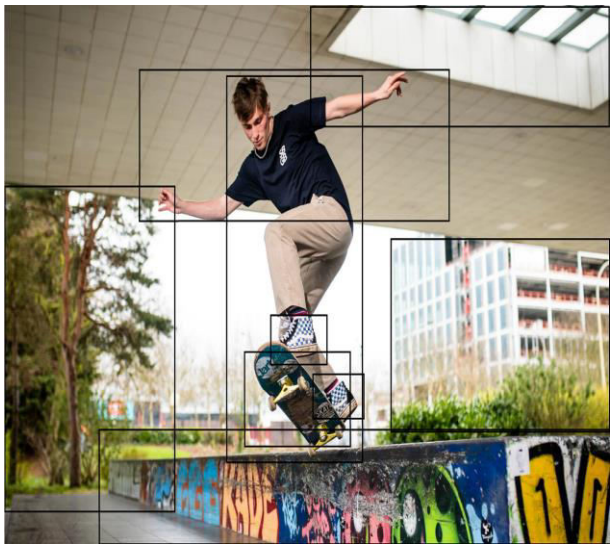
Abstract

We propose a Visual Question Answering System that enables deep image understanding through fine-grained analysis. Given an image and a question related to the image, the task is to provide an answer by understanding the image. Visual questions target different part of the image including the background of the image. As a result we need more thorough understanding of the image. In this work two approaches are being used, bottom-up and top-down attention mechanism. The bottom-up approach is based on Faster R-CNN (Regions with Convolution Neural Network), while the top-down mechanism is based on LSTM (Long Short Term Memory network) which is a special kind of RNN (Recurrent Neural Network). By using this approaches we are able obtain a better understanding of an image. We are able to solve the VQA challenge and also get better efficiency.

1.Introduction

The latest advances in computer vision have brought us closer to the point where traditional object recognition benchmarks (such as Image net) are considered "solvable." These advances, however, also prompt the question how we will move from object perception to visual understanding; that's , how we will extend today's recognition systems that provide us with "words" describing an image or an image area to frameworks which will deliver a more profound semantic representation of the image content. Since benchmarks have always been the most drive for the event of computer vision, some recent studies have proposed methods to guage our ability to develop such representations. These proposals include modelling relations between objects, visual Turing tests [4], and visual question answering.





In the human sensory system , attention are often focused volitionally by top-down signals determined by the presenttask (e.g., trying to find something), and automatically by bottom-up signals related to unexpected, novel or salient stimuli . during this paper we adopt similar terminology and ask attention mechanisms driven by nonvisual or task-specific context as 'top-down' and purely visual feed-forward attention mechanisms as 'bottom-up'.

Figure 1. Attention models work on CNN features relating to a uniform network of similarly – sized image regions(left). Our methodology empowers consideration regarding be determined at the degree of articles and other notable image regions(right).

Most conventional visual attention mechanisms used VQA are of the top-down variety. In the context of image-related problems, these mechanisms are usually trained to selectively participate in the output of one or more layers of a convolutional neural network (CNN). In any case, this methodology gives little consideration to how the image regions that are dependent upon consideration are resolved. As shown in the conceptual diagram in Figure 1, regardless of the image content, the generated input area corresponds to a consistent grid of neural receptive fields of the same size and shape. To get more human-like answers, objects and other salient image regions are a way more natural basis for attention

VQA seems to be a natural playground to develop approaches ready to perform basic "reasoning" about a picture . Recently, many studies have explored this direction by adding simple memory or attention-based components to VQA systems. While in theory, these approaches have the potential to perform simple

reasoning, it's not clear if they are doing actually reason, or if they are doing so in a human-comprehensible way.

In this paper we propose a combined bottom-up and topdown visual attention mechanism. The bottom-up mechanism proposes a set of salient image regions, with each region represented by a pooled convolutional feature vector. In fact, we use Faster R-CNN [32] to achieve bottom-up attention, which represents the natural expression of the bottom-up attention mechanism. The top-down mechanism uses task-specific context to predict the attention distribution on the image area. The feature vectors to be looked after are then calculated as the weighted average of the image features in all regions.

We evaluate the impact of mixing bottom-up and topdown attention.

2.Related Work

A large number of attention-based deep neural networks are proposed for VQA. Typically, these models are often characterized as top-down approaches, with context provided by a representation of the question within the case of VQA [11, 28, 44, 46, 49]. Attention is applied to the output of 1 or more layers of a CNN, by predicting a weighting for every spatial location within the CNN output. In any case, deciding the optimal number of image districts constantly requires an unwinnable exchange off among coarse and fine degrees of detail. In addition, the arbitrary positioning of the area relative to the image content may make it more difficult to detect objects that are poorly aligned with the area and to bind visual concepts associated with the same object. Relatively speaking, previous work rarely considers focusing on the salient image area. We realize that there are two papers. Jin et al.[18] use selective search [41] to find significant image areas, which will be filtered by a classifier, then resized and CNN encoded as input to the image caption model. During this work, instead of using hand-crafted or differentiable region proposals [41, 50, 17], we leverage Faster R-CNN [32], establishing a better link between vision and language tasks and up to date progress in object detection. We are ready to pre-train our region proposals on object detection datasets with this approach. Conceptually, the benefits should be almost like pre-training visual representations on ImageNet [34] and leveraging significantly larger cross-domain knowledge. We apply our method to VQA, establishing the broad applicability of our approach.

3.Approach

Given a picture I, our VQA model take as input a possibly variably sized set of k image features, V = {v1⋯., vk}, vi 2 RD,, such each image feature encodes a salient region of the image. The spatial image features V are often variously defined because the output of our bottom-up attention model, or, following standard practice, because the spatial output layer of a CNN. Our approach is to implement a bottom-up attention model in Section 3.1. In Section 3.2 we outline the Top-Down Attention model and in Section 3.3 we outline our VQA model. We note that for the top-down attention component, both models use simple one-pass attention mechanisms, as against the more complex schemes of recent models like stacked, multi-headed, or bidirectional attention that would even be applied.

3.1. Bottom-Up Attention Model

The definition of spatial image features V is not-specific. However, during this work we define spatial regions in terms of boxes and we implement bottom-up attention using Faster R-CNN. Faster R-CNN is an object detection model designed to spot instances of objects belonging to certain classes and localize them with bounding boxes. Other region proposal networks could even be trained as an attentive mechanism.

Faster R-CNN detects objects in two stages. the primary stage, described as a neighbourhood Proposal Network (RPN), predicts object proposals. A small-scale network is slid over features at an intermediate level of a CNN. At each spatial location, the network will predict class-agnostic objective scores and refine the bounding boxes for anchor boxes with multiple ratios and aspect ratios. Using greedy non-maximum suppression with an intersection-over-union (IoU) threshold, the highest box proposals are selected as input to the second stage. within the second stage, region of interest (RoI) pooling is employed to extract a small feature map for every box proposal. These feature maps are then arranged together as input to the ultimate layers of the CNN. the ultimate output of the model consists of a softmax distribution over class labels and class-specific bounding box refinements for every box proposal.

In this work, we use Faster R-CNN in conjunction with the ResNet-101 CNN. to get an output set of image features V to be used in VQA, we take the ultimate output of the model and perform non-maximum suppression for every object class using an IoU threshold. Then, we select all regions where the detection probability of all class exceeds the confidence threshold. for every selected region i, vi is defined because the

mean-pooled convolutional feature from this region, such the dimension D of the image feature vectors is 2048. When used in this way, Faster R-CNN effectively acts as a "hard" attention mechanism because a relatively small number of images bounding box features can only be selected from a large number of possible configurations.

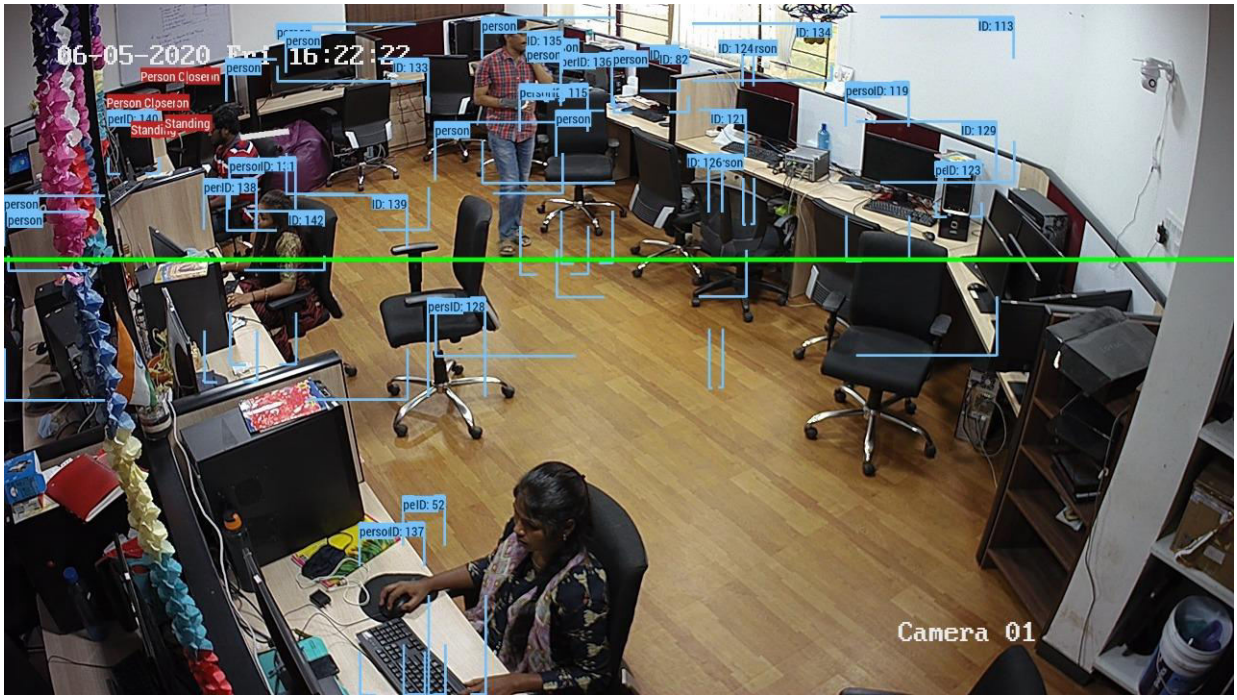


Figure 2. Example output from our bottom-up attentionmodel. Each box is labeled with an attribute classfollowed by an object class.

To pretrain the bottom-up attention model, we first initialize Faster R-CNN with ResNet-101 pretrained for classification on ImageNet. We then train on Visual Genome data. to assist the training of excellent feature representations, we add a further training output for predicting attribute classes (in addition to object classes). We concatenate the mean pooled convolutional feature vi with a learned embedding of the ground-truth object class, to predict attributes for region i and feed it to an additional output layer that defines the softmax distribution of each attribute class and "no attribute" class.

The original Faster R-CNN multi-task loss function contains four components, defined over the classification and bounding box regression outputs for both the RPN and therefore the final object class proposals respectively. We retain these components and add a further multi-class loss component to coach the attribute predictor. In Figure 2 we offer some samples of model output.

Our Faster R-CNN bottom-up attention model gives the output shown in Figure 2. Each bounding box is labelled with an attribute class followed by an object class. Note however, that in VQA we utilize only the feature vectors – not the anticipated labels.

3.2 Top-Down Attention LSTM

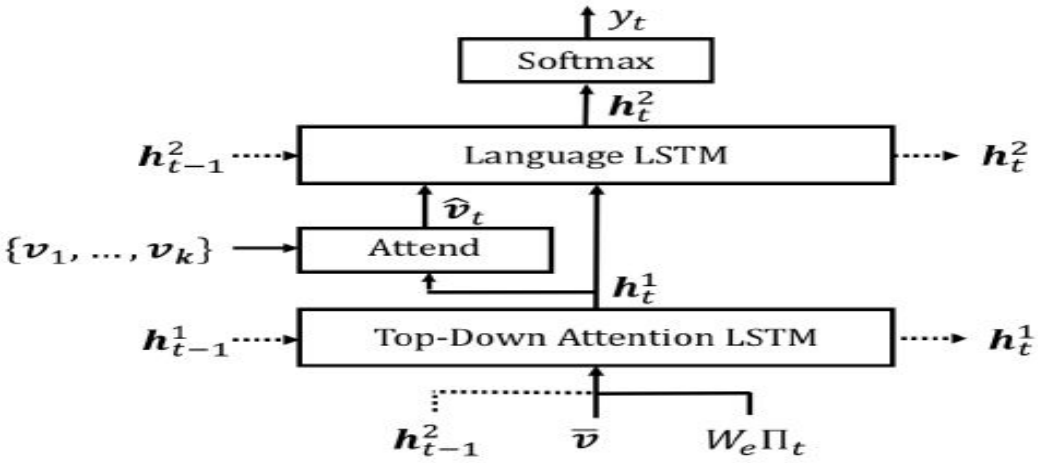


Figure 3. Overview of the top down attention model.

In the sections that follow we will refer to operation of the LSTM over one time step using the subsequent notation:

$h_t = \text{LSTM}(x_t, ht\text{-}1)$ (1)

where $x_t$is the LSTM input vector and $h_t$is the LSTM output vector. Here we have neglected the propagation of memory cells for notational convenience. We now describe the formulation of the LSTM input vector $x_t$and therefore the output vector $h_t$for every layer of the model.

Given the output $h^1{}_t$of the attention LSTM, at whenever step $t$we generate a normalized attention weight $\alpha_{i,t}$ for every kth image features v_i as follows:

$\alpha_{i,t} = w_a^T \tanh(W_{va} v_i + W_{ha} h_t^1)$ (2)

$\alpha_{i,t} = \text{softmax}(\alpha_t)$ (3)

where $Wva \in \mathbb{R}^{HxV}$, $Wha \in \mathbb{R}^{HxM}$ and $wa \in \mathbb{R}^{H}$ are learned parameters. The attended image feature used as input to the language LSTM is calculated as a convex combination of all input features:

$$\hat{v}_t = \sum_{i=1}^{K} \alpha_{i,t} v_i \quad (4)$$

3.3. VQA Model

Given a group of spatial image features V,our proposed VQA model also uses a 'soft' top-down attention mechanism which weighs each feature, using the question. As illustrated in Figure 3, the proposed model implements the well-known joint multimodal embedding of the question and the image, followed by a prediction of regression of scores over a set of candidate answers. This approach has been the idea of various previous models.

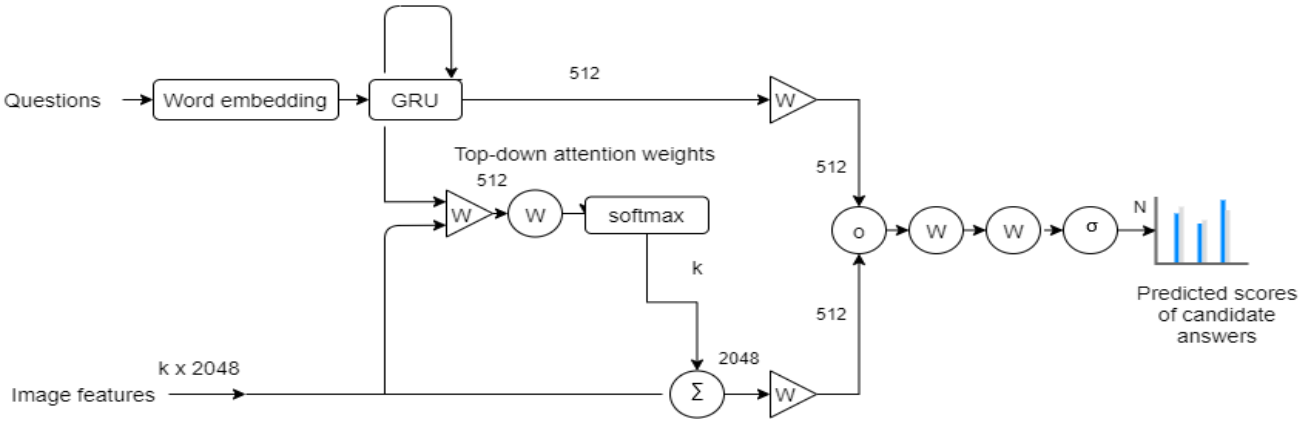


Figure 4. Overview of the proposed VQA model.

The learned non-linear transformations within the network are implemented with gated hyperbolic tangent activations. These are a special case of highway networks that have shown a strong empirical advantage over traditional ReLU or tanh layers. Each of our 'gated tanh' layers implements a function $fa: x \in \mathbb{R}^m \rightarrow y \in \mathbb{R}^n$ with parameters $a = \{W, W', b, b'\}$ defined as follows:

$$\tilde{y} = \tanh(Wx + b) \quad (5)$$

$$g = \sigma(W' x + b') \quad (6)$$

$$y = \tilde{y} \circ g \quad (7)$$

where σ is the sigmoid activation function, $W, W' \in \mathbb{R}^{mxn}$ are learned weights, $b, b' \in \mathbb{R}^{n}$ are learned biases, and $\circ$ is the Hadamard (element-wise) product. The vector g acts as a gate to activate $\tilde{y}$ in the middle.

Our proposed method first encodes each problem as the hidden state q of a gated loop unit (GRU), and each input word is represented using a learning word embedding. Similar to Equation 2, given the output q of the GRU, we generate an unnormalized attention weight ai for each of the k image features vias follows:

$$\alpha_i = w_a^T f_a([v_i, q]) \quad (8)7$$

where $w_a^T$is a learned parameter vector. Equation 3 and Equation 4 (neglecting subscripts t) are used to calculate the normalized attention weight and the attended image feature $\hat{v}$. The distribution of the possible output response y is given by:

$$h = fq(q) \circ fv(\hat{v}) \quad (9)$$

$$p(y) = \sigma(Wo\, fo(h)) \quad (10)$$

Where h may be a joint representation of the question and therefore the image, and Wo $\in \mathbb{R}^{|\Sigma| x M}$are learned weights.

thanks to space constraints, some important aspects of our VQA approach aren't detailed here. For full specifics of the VQA model including an in depth exploration of architectures and hyperparameters, ask Teney et al.

4.Evaluation

We evaluated the following supremacy of our extricated structural segment representation and symbolic execution engine. Firstly, our model can grasp from a small amount of training data and outmatch the latest state-of-the-art techniques while correctly retrieving the dormant programs. Secondly, our model popularizes well to other asking styles, characteristics mixture ,and visual surroundings.

4.1 Dataset

In this part, we have issued an inspection of the questions and answers in the VQA train dataset. To obtain an apprehension of the types of questions queried and answers supplied, we imagine the diffusion of question types and answers. We also inspected how frequently the questions may be answered in the absence of the picture using just basic details. At last, we examined whether the data present in an image is sufficient to answer the questions. The dataset includes 614,163 questions and 7,984,119 answers (including answers without glancing at the picture) for 204,721 images from the dataset and 150,000 questions with 1,950,000 answers for 50, 000 unique segment.

To evaluate our suggested VQA model, we use the latest launched VQA v2.0 dataset, which tries to reduce the productiveness of understanding dataset before by comparing the answers to every question. The dataset, contains 1.1M questions with 11.1M answers relating to MSCOCO images. We mounted typical question text preprocessing and tokenization. Questions are snipped to a maximum of 14 terms for computational capability. The group of listed answers is limited to right answers in the training set that is visible more than 8 times, prompting in a vocabulary size of 3,129.

Our VQA test server proposals are trained on the validation and training sets plus extra questions and answers from Visual Genome. To assess the quality of an answer, we describe accuracies using the typical VQA metric, which takes into consideration the rare disapproval among annotators for the ground real answers.

The Visual Genome dataset is used for data augmentation and in pre-training ourbottom-up attention model, when training our VQA structure.For pre-training the bottom-up attention model, we use only the attribute and object data. We stored 5K pictures for authentication, and 5K images for later trials, considering the rest 98K images as training data. We made sure that any images found in any other datasets are found in the same lop in both datasets.

As the attribute and object annotations consist of openly interpreted threads, rather than classes, we carried large-scale filtering and cleaning of the training data. First from500 attribute classes and 2,000 object classes, we physically erased abstract classes that gave bad detection execution in earlier tests. Our last training set comprises of 400 attribute classes and 1,600 object classes.

Note that we did not combined or erased intersecting classes (e.g. ‘guy’, ‘man’, ‘person’), classes with both plural and singular types (e.g. ‘trees’, ‘tree’) and classes that are hard to accurately localize (e.g., ‘buildings’, ‘grass’, ‘sky’).When training the VQA model, we increased the VQAv2.0 training data with Visual Genome

question and answer sets gave the right answer is visible in model's answer vocabulary. This constitutes about 30% of the present data, or 485K questions.

To justify the influence of bottom-up attention, in VQA experiments we assess our full structure in case of previous efforts as well as a withdrawn baseline. In every scene, the baseline, takes a ResNet CNN pretrained on ImageNet to cipher each picture in position of the bottom-up attention mechanism. In VQA experiments, we cipher the rescaled input image with ResNet-200. In another experiments we tested the impact of differing the scale of the spatial output from its standard scale of 14_14, to 7_7 (using bilinear interpolation) and 1_1.

4.2 Questions

The next evaluation is in terms of the questions framed and presented. In the composition of questions createdin the English language, we can bunch questions intovarious types on the basis of words that begin the question. Surprisingly, the arrangement of questions is very much alike for both symbolic views and real images. This aids to represent that the kind of questions obtained by the symbolic views is similar to those obtained by the real images.

There subsists a interesting options ofquestion types, including "Does the. . .", "Is there. . .", "Howmany. . .", and "What is. . .". A certainly unusual type of question is "What is. . ." questions, as they consist of a large pool of feasible answers. In terms of lengths we saw that many of the questions range from four to ten words only.

4.3 Answers

Next is the evaluation of the answers provided and in many typical answers we can see that a number of question categories, such as "Are. . . ", "Is the. . . ", and "Does. . . " are particularly answered using "no" and "yes" as responses. Further questions such as "What type. . . " and "What is. . . " have an abundant variety of answers. Additional question types such as "Which. . . " or "What color. . . " have more specific answers, such as "right" and "left" or colors.

The lengths of almost all answers are in range of one word, with the diffusion of responses containing one, two, or three words,respectively being 90.51%, 5.89%, and 2.49% for symbolic viewsand 89.32%, 6.91%, and 2.74% for real pictures. The shortness of responses is not a new case, as the questions are likely to obtain particular data from the pictures.

The shortness of our responses makes instinctive evaluation possible. While it may be enticing to trust the shortness ofthe answers makes the problem simpler, recollect that they are man-provided unlimited answers to

unlimited questions. The questions mostly need high level reasoning to reach at these misleading easy responses.

There are at present 3,770 for symbolic views and 23,234 abstract one-word responses in ourdataset for real images. Numerous questions are answered using either “no” or “yes” (or at times with “maybe”)– 40.66% and 38.37% of the questions on symbolic viewsand real images respectively. Out of these ‘yes/no’ questions, there is a partiality towards “yes” – 58.83% and 55.86% of ‘yes/no’ responses are “yes” for symbolic views and real images. Question types such as “How many. . . ” are responded using numbers – 12.31% and 14.48% of the questions on real images and symbolic views are ‘number’ questions.“2” is the utmost favored response among the ‘number’ questions, taking up 39.85% for symbolic viewsand 26.04% of the ‘number’ responses for real images.

In terms of subject confidence, when the subjects answered the questions, we had put forth “Do you think you were able to answer thequestion correctly?” to make them sure about their responses and ininter-human agreement we thought of Does the self-awareness of belief with respect to the responses agrees between subjects? As predicted, the acceptance among subjects rises with confidence.

Nevertheless, even if all are confident the responses may still differ. This is not new as some responses may differ, yet have very much same meaning, such as “joyful” and “happy”. There is a notable inter-human agreement in the responses for both symbolic views (87.49%) and real images(83.30%).

4.4 Commonsense Knowledge

Is the Picture essential? Clearly, some questions can sometimes be answered properly using basic commonsense knowledge only and there is no need for an image. For “yes/no” questions, the man subjects answers better than prospected. For additional questions, human are only right about21% of the time. This reveals that knowing the visual knowledge is essential to VQA and that commonsense knowledge alone is not enough.

To find the actual difference in responses provided with and without pictures, we have revealed the diffusion of answers for different question types. The distributionof numbers, colors, and even “yes/no” answers is shockingly varies for responses with and without pictures. Which Questions NeedBasic Common Sense? In view to finding questions that needed commonsense reasoning to responses, we performed two experiments asking subjects– (i) whether or not a question required information outside to the picture, and (ii) the smallest age group people thatcould answer the question – adult (18+), teenager (13-17),older child (9-12),younger child (5-8), toddler (3-4). Every question was given to 10 subjects. We claimed that for47.43% of question 3 or

more subjects chosen 'yes' to commonsense,(18.14%: 6 or more). The following is the distribution of responses that were recorded: younger child:39.7%, older child: 28.4%, toddler: 15.3%, teenager: 11.2%, adult: 5.5%.

4.5 Questions vs Captions

Do general captions give all information toanswer the questions? Yes, But however, the precisions are greatly lesser than when subjects are shown the real picture. This shows that in order to answer the questions properly, greater image knowledge is essential. In fact, we discovered that the diffusion of verbs, adjectives and nouns specified in captions is numerically unorthodox from those specified in our answers + questions for both symbolic views and real images.

5.Results / Snapshot

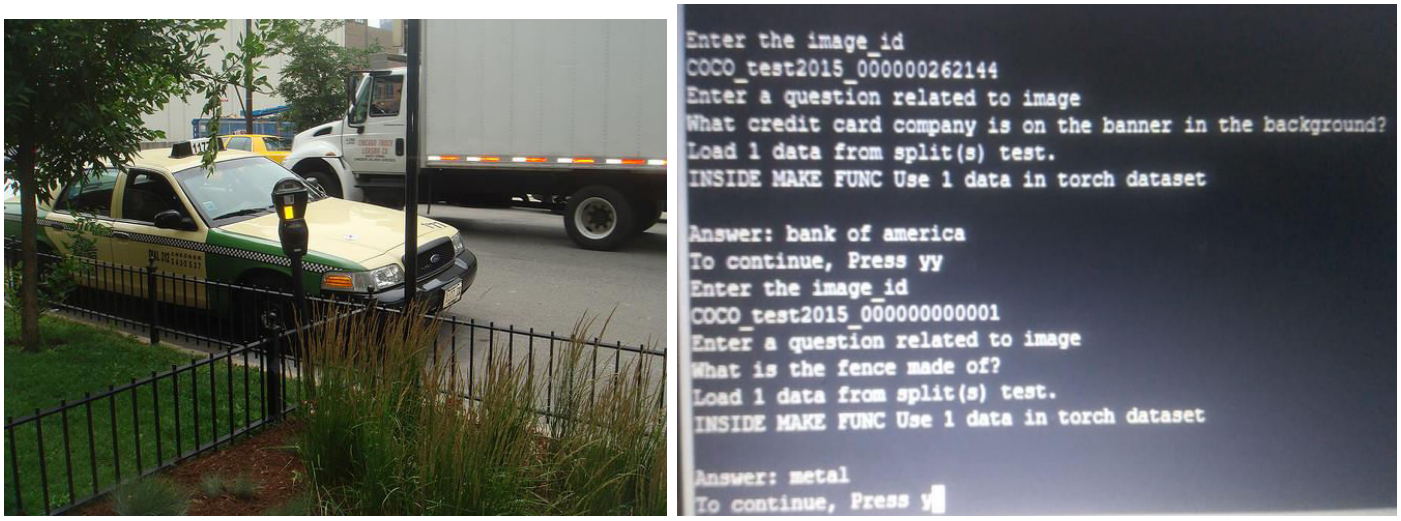


Figure 5. Image and related question and answer.

6.Conclusions

We have introduced an experimental top-down and bottom-up collaborated visual question answering system. Our visual attention mechanism enables the value to be calculated more naturally and with greater depth at the level of objects and other salient regions of the image. Putting this approach to the visual question answering system, we have achieved futuristic and up to the minute results in our tasks, while refining the understandability of the desired output.

The correctness of the answers is as expected and we have got the results as desired. There can be a variety of questions that we can ask and will get the appropriate answers accordingly. The challenge of handling the large dataset was also eased by the pickle program that really helped us to complete this project and with such a great precision.

7.FutureEnhancements

At a bigger extent, our work more firmly fuses tasks implying optical and semantic understanding with current advancement in object detection. While this hints various developments for future research in the area where the precision can somewhat be upgraded, the instant profits of our approach is aquired by simply restoring pre-trained CNN features with pre-trained bottom-up attention attributes.

References

P. Anderson, B. Fernando, M. Johnson, and S. Gould.SPICE: Semantic propositional image caption evaluation.

S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L.Zitnick, and D. Parikh. VQA: Visual Question Answering.

T. J. Buschman and E. K. Miller. Top-down versus bottomupcontrol of attention in the prefrontal and posterior parietalcortices. *Science*,

X. Chen, T.-Y. L. Hao Fang, R. Vedantam, S. Gupta,P. Dollar, and C. L. Zitnick. Microsoft COCO captions:Data collection and evaluation server. *arXiv preprintarXiv*

K. Cho, B. van Merrienboer, C. Gulcehre, F. Bougares,H. Schwenk, and Y. Bengio. Learning phrase representationsusing RNN encoder-decoder for statistical machine translation.

M. Corbetta and G. L. Shulman. Control of goal-directedand stimulus-driven attention in the brain. *Nature Reviews*

Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier. Languagemodeling with gated convolutional networks. *arXiv preprintarXiv:1612.08083*, 2016.