CS381V Final Project Report
Visual Question Answering using Natural Language Object Retrieval and Saliency Cues

Aishwarya Padmakumar
aish@cs.utexas.edu

Akanksha Saran
asaran@cs.utexas.edu

Abstract

The goal of this project is to examine the effect of explicitly providing two additional sources of information to a Visual Question Answering system apart from the original image based on which a question is asked. The first is a bounding box obtained by performing natural language query based object retrieval, using the question as the query. This is expected to identify a region of interest in the image very likely to contain the answer to the question. Such a system allows for the use of synonyms and compositional descriptions in the question still corresponding to a particular object or region of interest in the image. The second source of information is a bounding box extracted from the saliency map of the image- a region in the image that humans typically find interesting and are hence likely to have asked questions about. We provide these sources of information (CNN features for the bounding box selected by the saliency map, CNN features for the bounding box selected by the natural language object retrieval pipeline and the question encoded by an LSTM) as inputs to a neural network and train a loss function to optimize for the task of VQA on the MS-COCO VQA dataset. We analyze various choices of layers for combining embeddings of the various inputs and find that the best performing system, which takes 4 inputs (original image, saliency region, Natural Language Object Retrieval region and embedded question) beats the baseline by 0.45 percentage points.

1. Introduction

With the recent advances in image recognition tasks, Visual Question Answering is now emerging as a new challenge for the computer vision community. It requires combining multi-modal knowledge, use of common-sense knowledge and a deeper understanding of vision and language than many other AI tasks [2] which makes it compelling as an AI complete problem. Recently, there have been many attempts towards solving this problem which we outline in Section 2.

While some of these aim to attend to specific parts of the image as different words of the question are read, they do not attempt to ground the question as a whole to objects of importance in the image. Hence for a question such as *What colour is the shirt of the man near the cat?*, a good attention model would focus on shirts in the image when reading *shirt*, men in the picture when reading *man* and on a cat when reading *cat*, it does not necessarily perform compositional reasoning to understand that *man near the cat* refers to a specific man in the image. We extend the work by Hu *et al.* [6] to make use of such reasoning to improve visual question answering. In [6], given a natural language query such as *man in the middle with blue shirt and blue shorts*, the goal is to retrieve a single bounding box or multiple bounding boxes (ranked by relevance with respect to the question) around the man satisfying this description (figure 1). We propose to combine this with a standard VQA pipeline in the following manner - instead of having a single CNN which reads the input image, we add another CNN to which the top ranked bounding box provided by the natural language object retrieval pipeline is provided as input. We also use a model which combines the CNN features of top three bounding boxes (selected by the natural language object retrieval pipeline) weighted by the scores corresponding to each of these boxes.

We believe we are the first to incorporate explicit visual grounding of the question when performing VQA. We have also not found any relevant work that performs the simpler extension of providing detections of objects mentioned in the question. We believe that retrieving objects using the entire question as a query is likely to be more useful than the simpler alternative because of two reasons -

- Typically, learned object detectors will be associated with a single word. However synonyms (pony instead of horse), hypernyms (animal instead of cat) or hyponyms (spaniel instead of dog) of this word may be used in the question and we may not realize that an available detector is applicable.

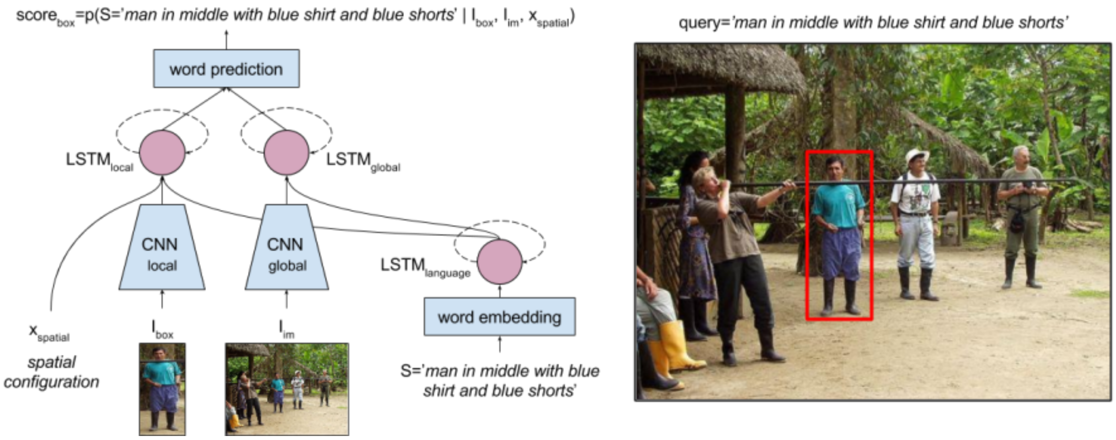

score$_{box}$=p(S='man in middle with blue shirt and blue shorts' | I$_{box}$, I$_{im}$, x$_{spatial}$)
word prediction
LSTM$_{local}$
LSTM$_{global}$
CNN local
CNN global
LSTM$_{language}$
word embedding
x$_{spatial}$
spatial configuration
I$_{box}$
I$_{im}$
S='man in middle with blue shirt and blue shorts'
query='man in middle with blue shirt and blue shorts'

Figure 1: Natural Language Object Retrieval pipeline [6].

- Such a method would still not understand a composite phrase such as *man near the cat* and hence probably not provide much more information than spatial attention.

Similar to attention models, this identifies regions in the image that are highly relevant to the question. In addition to this, we wish to use the most salient region in the image. Such salient regions can be extracted by thresholding saliency maps [9] to provide another notion of importance - which regions are generally likely to be of interest to humans, and hence likely to provide answers to questions asked by humans.

The remaining outline of the paper is as follows. We discuss related work on VQA in Section 2 and background work related to Natural Language Object Retreival and Saliency in Section 3. We then explain our approach and describe the different models based on our idea in Section 4. We go on to show qualitative and quantitative results for our different models and compare them to the baseline approach of [2] in Section 5. Finally we conclude with some insights in Section 6 and discuss potential future work in Section 7.

2. Related work

2.1. Recent work with results on the VQA dataset

Since the release of the VQA challenge dataset [2], there have been a number of works attempting to solve this problem. Some of these results can be seen in Table 1. LSTMIMG [2] is the basic method suggested in the original paper which released the dataset. Most of these techniques contain a CNN and LSTM at the heart of their architecture. A large number of them, including Ask, attend and Answer [20], Compositional Memory [8], ABC-CNN [3] and Stacked Attention Network (SAN) [21] are architectural variants designed to enable the system to pay attention to specific parts of the image as it processes each word in the question. DPPNet is also similar in goal but uses a different type of recurrent neural network - a GRU instead of an LSTM. A related technique that performs very well is Word + Region Selection [17] which maps text in the question and the image to a multimodal space to identify regions of interest in the image. In contrast to these approaches, we explicitly identify the region of interest and provide it as an additional input along with saliency information for the entire image. Thus we have two priors - natural language object retrieval provides a region of high relevance to the question whereas saliency provides information of important objects in the scene regardless of the question.

There have also been attempts that do not attempt to add an attention mechanism. NMN + LSTM [1] is a model that attempts to use classical natural language processing techniques to determine the specific type of computer vision task required to answer the question and calls upon an appropriate neural module. iBOWIMG [24] is a surprisingly simple technqiue that eschews a recurrent neural network for a simple bag of words representation and interestingly outperforms even many models using attention. ACK [19] attempts to incorporate knowledge from external knowledge sources to enable the answering of commonsense questions that cannot be directly answered using the information in the image.

The most similar work on this dataset is the very recent work of Ilievsky *et al*. [7] (FDA). They use off-the-shelf object detectors to identify objects present in the question and supply objects whose labels have a high similarity, according to word2vec [14], to an LSTM and finally use its learned representation along with an encoding of the question to answer the question. While the form of supervision is conceptually similar to ours, we believe that using natural language object retrieval will better handle problems such as synonymy, hypernymy and hyponymy, as well as being able to resolve compositinal expressions to a single object.

Models	Accuracy (%)
ABC-CNN [3]	48.38
Compositional Memory [8]	52.68
LSTMIMG [2]	54.06
NMN + LSTM [1]	55.10
iBOWIMG [24]	55.89
ACK [19]	55.98
DPPnet [15]	57.36
Ask, Attend and Answer [20]	57.99
SAN [21]	58.7
Word + Region Selection [17]	62.44
FDA [7]	**64.18**

Table 1: Existing approaches for VQA which report performance on the MSCOCO-VQA dataset [2]

2.2. Other approaches and datasets

[25] introduces a new dataset called 'Visual7W' based on the MS COCO dataset for visual question answering using both multiple-choice textual answers (telling) and visual answers (pointing). They use a novel LSTM architecture which models attention towards local regions in the image based on the words in the question. They show that their dataset has the largest gap between a baseline performance versus human performance, as compared to other VQA related datasets. They thus make the case for their dataset being the most challenging compared to other pre-existing ones in the domain. Their dataset also has bounding box annotations in terms of visual answers, however they do not use any level of region proposal or object detection to show results on their dataset.

[22] also introduces a new dataset called 'Visual Madlibs' with 10,738 images (subset of the human-centric MS COCO dataset) and 360,001 targeted descriptions. The descriptions are collected by using fill-in-the-blank templates to human annotators. Fill-in-the-blank questions can be targeted to collect descriptions about people and objects, their appearances, activities, and interactions, as well as descriptions of the general scene or the broader emotional, spatial, or temporal context of an image. This paper does not really pose the task being solved as visual question answering in the traditional sense but is more related to captioning. However they go a step beyond image captioning because their fill-in-the-blank templates focus on targeted natural language descriptions of image content that go beyond describing which objects are in the image, and beyond generic descriptions of the whole image. The authors create a multiple-choice QA task and use the filled templates as answers. They use a CNN+LSTM network similar to [2] to evaluate on the multiple-choice QA task. They also use a RCNN based object detector for the attribute related questions and find that they do as well as using the ground truth bounding boxes available with the COCO dataset. However, they do not attempt to perform compositional natural language understanding.

[12] is another paper which talks about the collection of the DAQUAR dataset for the task of visual question answering for real-world indoor scenes from the NYU dataset. The authors do not show any baseline results on this dataset but support the use of the WUPS score as a good metric of evaluation. The authors also present an approach [11] which combines semantic parsing and image segmentation. Their approach is notable as one of the first attempts at image QA, but it has a number of limitations. First, a human-defined possible set of predicates are very dataset-specific. To obtain the predicates, their algorithm also depends on the accuracy of the image segmentation algorithm and image depth information. Second, their model needs to compute all possible spatial relations in the training images. Even though the model limits this to the nearest neighbors of the test images, it could still be an expensive operation in larger datasets. [13] propose a deep architecture called Neural-Image-QA, where the image is analyzed via a CNN and the question together with the visual representation is fed into a LSTM network. CNN and LSTM are trained jointly and end-to-end starting from words and pixels. This paper reports performance on the DAQUAR dataset but their pipeline is essentially the same as the baseline presented in the VQA challenge paper [2].

[16] propose to use neural networks and visual semantic embeddings, without intermediate stages such as object detection and image segmentation, to predict answers to simple questions about images.

[10] use a model consisting of three CNNs: one image CNN to encode the image content, one sentence CNN to compose the words of the question, and one multimodal convolution layer to learn their joint representation for the classification in the space of candidate answer words. They report their results on the DAQUAR dataset and the COCO-QA dataset.

[4] In this paper, the authors present the mQA model, which is able to answer questions about the content of an image. The answer can be a sentence, a phrase or a single word. Their model contains four components: a Long Short-Term Memory (LSTM) to extract the question representation, a Convolutional Neural Network (CNN) to extract the visual representation, an LSTM for storing the lin-

guistic context in an answer, and a fusing component to combine the information from the first three components and generate the answer. They also construct a Freestyle Multilingual Image Question Answering (FM-IQA) dataset to train and evaluate their mQA model. It contains over 150,000 images and 310,000 freestyle Chinese question-answer pairs and their English translations.

3. Background

3.1. Natural Language Object Retrieval

Natural language object retrieval (NLOR) [6] addresses the problem of localizing a target object within a given image based on a natural language query describing the object. This work uses as input the original image, natural language queries describing the target object and bounding box proposals obtained from Edge Box [26] and outputs scores for each input bounding box corresponding to the likelihood of the target object being contained within it. Edge box [26] generates about 100 object bounding box proposals per image using edges. The intuition behind using edges to generate bounding boxes is that the number of contours that are wholly contained in a bounding box is indicative of the likelihood of the box containing an object.

NLOR [6] uses considers all the bounding boxes generated by Edge Box [26] as equally likely to begin with, and ranks them based on the query. In particular, [6] leverages both spatial information about objects within the scene and global scene context. A novel Spatial Context Recurrent ConvNet (SCRC) model is proposed as a scoring function on candidate boxes for object retrieval, integrating spatial configurations and global scene-level contextual information into the network. Processing of query text, local image descriptors, spatial configurations and global context features through a recurrent network, it outputs the probability of the query text conditioned on each candidate box as a score for the box. An illustrative overview of this approach is shown in Fig.1

3.2. Saliency

For obtaining a more general saliency signal, we use the work of Liu *et al*. [9]. Their goal is to separate a single, salient object in the image from the background, by casting salient object detection as a binary labelling problem. Saliency is scored using a number of cues, such as contrast over a Gaussian pyramid, a center-surround histogram to detect whether a patch is distinctive with respect to its surrounding patches, and the spatial variance of colour. These are combined using a CRF to obtain a saliency score between 0 and 1 for every pixel in the image. The combination of the different types of features helps overcome some limitations of each individual feature, such as the fact that contrast fails to capture the interior of a salient object, center-surround histograms may fire on the background, and the low precision fo colour-based features.

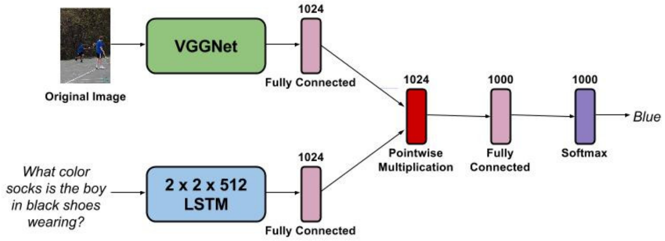


Figure 2: VQA Baseline Model

This work is evaluated in terms of its ability to identify objects marked as salient by humans, which makes it relevant to the task of Visual Question Answering, as the hope is that such an object would be most likely to attract the attention of a human questioner, and hence contain the answer to the question. Also, although the authors aim to identify a single, salient object in an image, the technique seems directly applicable even in the case of multiple salient objects being present.

3.3. VQA Baseline

We use the VQA pipeline suggested in [2] as our baseline for comparison (*baseline*), and to extend with our additional inputs. This model can be seen in figure 2. The question is embedded using an LSTM [5] into a vector space, and the image by a convolutional neural network - VGGNet [18] pre-trained on ImageNet, which are combined using pointwise multiplication. The final answer is obtained by a classification over the 1000 most common answers in the dataset.

4. Approach

To obtain regions of interest using NLOR, we first obtain 100 candidate object proposals for the image using EdgeBox [26]. These are then re-ranked using the NLOR pipeline from Hu *et al*. [6]. We then obtain crops of the image corresponding to the top three bounding boxes according to this ranking. This is done as a separate pre-processing step as NLOR needs to examine the candidate object proposals in sequence, making it difficult to directly integrate with a VQA system.

To obtain regions of interest according to saliency, we apply the method of Liu *et al*. [9] to obtain a saliency map over the image. We obtain the tightest crop of the image containing the largest connected component obtained when this saliency map has been thresholded. An alternative approach would have been to add the saliency map as an additional channel to the image, but this would have required re-training of the VGGNet on the whole of ImageNet.

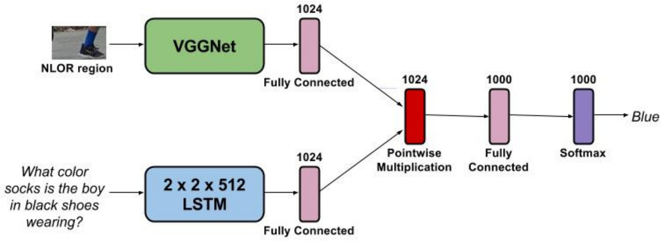


Figure 3: *Model NLOR* - Top region from NLOR replaces original image

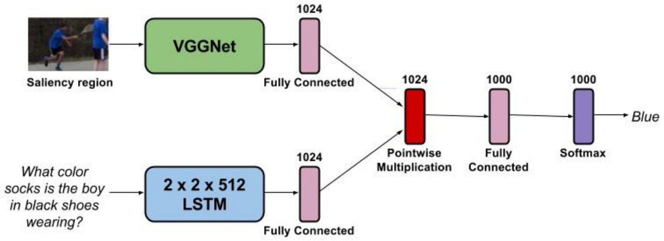


Figure 4: *Model Sal* - Connected component from saliency map replaces original image

4.1. Models using only regions of interest

In our first two models, we retain the model structure of the VQA baseline model but instead of providing the entire image, we provide just a region of interest. In *Model NLOR* (figure 4), the original image is replaced with the crop corresponding to the top ranked bounding box from NLOR. In *Model Sal* (figure **??**), the original image is replaced with the crop containing the largest connected component in the thresholded saliency map. In both these models, the question and image embeddings are combined by pointwise multiplication as in the baseline. If either of these models outperform the baseline, it would indicate that the regions of interest are fairly accurate at capturing the answer and hence, focusing on this region prevents focus on unimportant clutter in the image.

4.2. Models using original image and regions of interest

In these models, we wish to provide both the original image and regions of interest from NLOR and/or saliency, allowing the network to extract information from both. The first of these is *Model NLOR_M*, in which the region from NLOR is embedded using a VGGNet pre-trained on ImageNet, as is the original image, and the embeddings of these and the question are combined by pointwise multiplication of all three vectors. This can be seen in figure 5. Further, since we were not convinced that pointwise multiplication was the best method to combine three embeddings, we also attempted two other methods. One of these was simple concatenation of the three vectors (*Model NLOR_C* - figure 6). The other was to first separately combine the embeddings of each image with the question by pointwise multiplication, and then concatenate the two resultant vectors (*Model NLOR_MC*), which can be seen in figure 7.

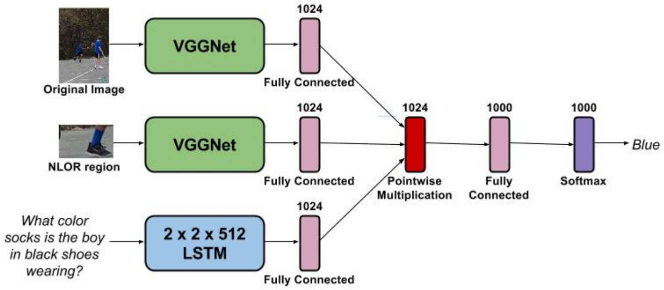


Figure 5: *Model NLOR_M* - Original image, region from NLOR and question combined by pointwise muliplication

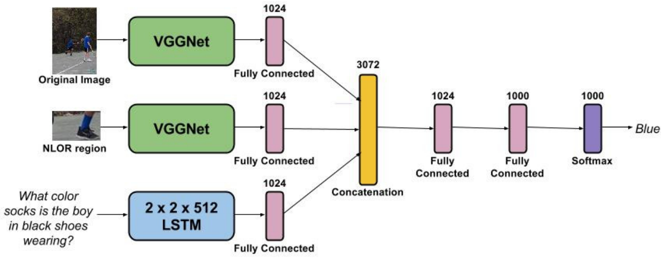


Figure 6: *Model NLOR_C* - Original image, region from NLOR and question combined by concatenation

As a direct comparison, we also attempted providing the region from saliency instead of the region from NLOR. Since the pointwise multiplication model was found to perform the best from the above three, we did not explore the other variants for this case. This model, *Model Sal_M* can be seen in figure 8.

Finally, since we believed that NLOR and saliency capture complementary types of information, providing both regions should be helpful for the overall task. To test this, we create a pipeline called *Model NLOR_Sal* (figure 9) where the original image, the region from NLOR, and the region from saliency are first embedded using individual CNNs and all these embeddings are combined with that of the question using pointwise multiplication. We did not try the other variants because of poor results in the NLOR case.

4.3. Models using multiple regions of interest from NLOR

Experimental results showed that the models using only regions from NLOR and the original image did not perform very well. We hypothesized that this could be due to the top bounding box from NLOR not necessarily containing the

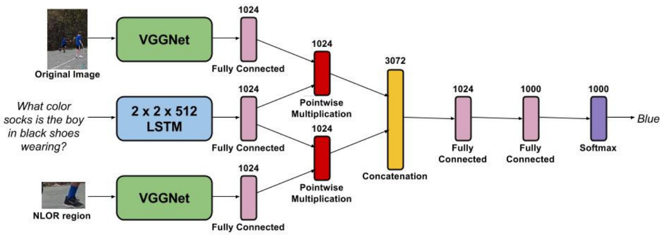


Figure 7: *Model NLOR_MC* - Original image and region from NLOR combined separately with question by pointwise multiplication, followed by concatenation

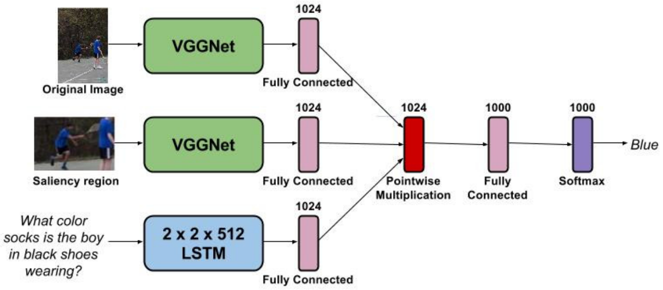


Figure 8: *Model Sal_M* - Original image, region from saliency and question combined by pointwise muliplication

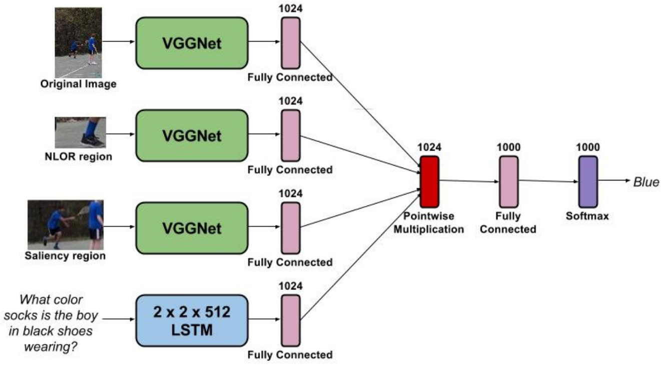


Figure 9: *Model NLOR_Sal* - Original image, region from saliency and question combined by pointwise muliplication

answer. We wished to explore whether using a larger number of regions from NLOR would help. The first of these models was *Model NLOR_3B* in which we embed the top three regions from NLOR using individual CNNs and combine them by a weighted linear combination, using scores from the NLOR system to weight the regions. This vector is then combined with the embedding of the original image and that of the question by pointwise multiplication. This can be seen in figure 10

We also attempt to incorporate saliency into this model by adding an embedding of the region from saliency as another factor combined in the pointwise multiplication *Model NLOR_3B_Sal*, which can be seen in figure 11

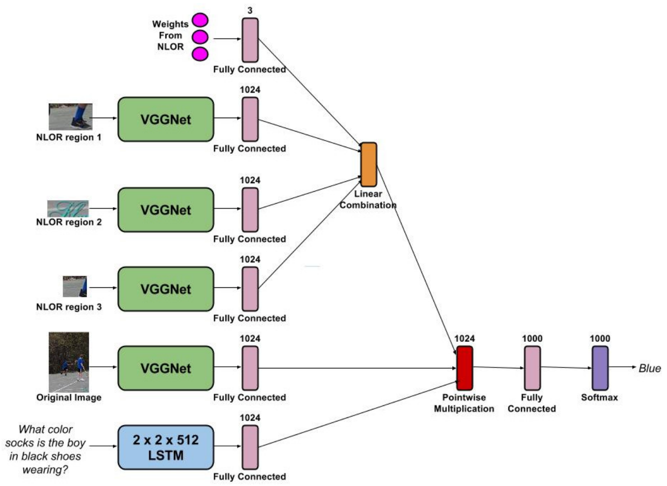


Figure 10: *Model NLOR_3B* - 3 regions from NLOR weighted by scores from NLOR, combined with original image and question by pointwise multiplication

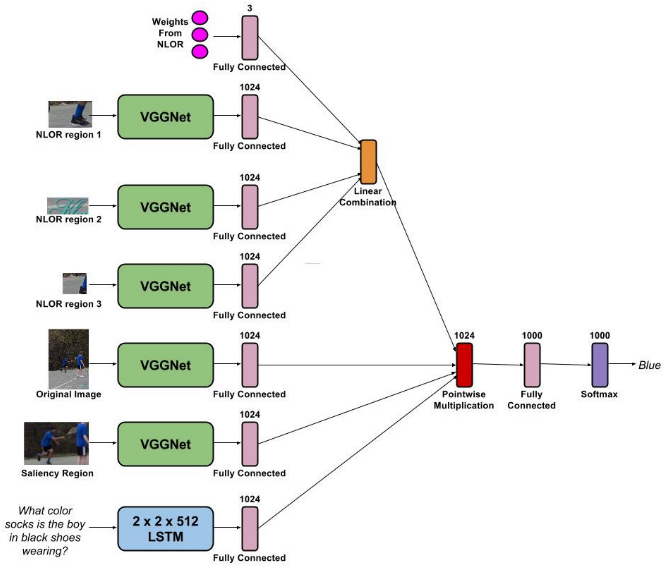


Figure 11: *Model NLOR_3B_Sal* - 3 regions from NLOR weighted by scores from NLOR, combined with original image, region from saliency and question by pointwise multiplication

5. Experimental Results

We illustrate qualitative and quantitative results pertaining to our models on VQA. We also show qualitative results for regions selected from NLOR and saliency cues.

5.1. Qualitative Results

We show illustrations for regions picked by NLOR, saliency cues and answers generated for VQA below.

5.1.1 NLOR

Examples of different regions selected by NLOR based on the questions asked for a given image are shown in Fig. 12, 13 and 14. We find that selected bounding boxes are sensible when the object about which the question is asked is mentioned clearly in the question. We also show some examples where NLOR bounding boxes do not make sense with respect to the question. More detailed explanation is available as part of the caption for the images

5.1.2 Saliency

Some examples of the regions identified by saliency can be seen in figures 15, 16, 17, 18, 19, and 20. There is also an analysis of which of the questions provided for the image would be answerable given the saliency region. Although our initial hypothesis was that saliency would identify regions of the image that are of interest to humans, and hence would be more likely to contain the answers to questions they ask, we find that people do seem to ask questions on other details as well, including questions based on the background.

5.1.3 VQA

We show qualitative results for four of our models in Fig. 21, 22, 23, 24, 25, 26 and 27. In each Figure the original image, the top 3 NLOR bounding boxes (ranked in order of relevance), the inherently salient region of the image, the three corresponding questions for that image, and the answers predicted by various models. Detailed explanation is provided below each image discussing the success and failures of the different methods.

5.2. Quantitative Results

Below we describe quantitative results of our various models in terms of the overall performance on all questions, a break down of how many question types our models outperform on compared to the baseline. We also report results on different answer types.

5.2.1 Overall performance of proposed models

We experimented with various models as described in Section 4. We can see the overall performance of all our models in comparison to the baseline approach of [2]. Our results are reported on the validation data. Most of the state of the art models report results on the test data and not on the validation data hence we don not compare with those here. We also analyze the results on the validation data over multiple iterations and report the number of iterations which gives the best result for each model on the validation data. We see that the only model which performs better than the baseline (54.51% versus 54.06%) is the one which uses as input VGGNet features of the original image, the top bounding box from NLOR, the salient region and the embedded question (Model NLOR_Sal). One reason we believe this could be happening is that this model incorporates what one would expect to intuitively be the most diverse and relevant information. Among the models which only use NLOR bounding box(es) and no saliency information, we see that the model which use the top NLOR bounding box, the original image and the question (combined using point wise multiplication) does the best (Model NLOR_M) with 53.97%. The fact that the model using top three bounding boxes doesn't do as well (53.62%) could be because the number of parameters of that model increase three times and the amount of data in terms of image-question pairs is insufficient to obtain good parameter estimates for this complex a model.

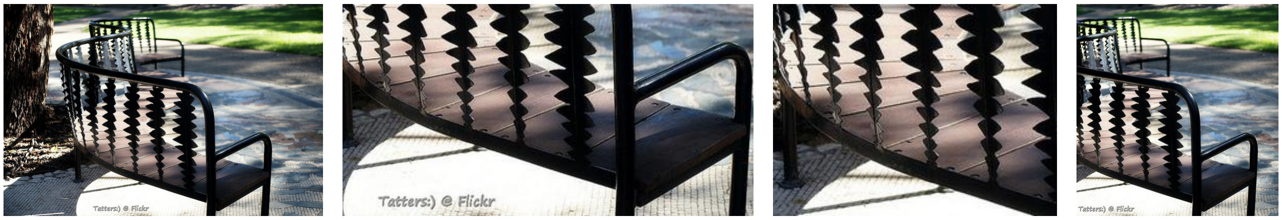
We tuned the learning rate parameter coarsely on all our models. We did not see any improvement in overall performance by coarsely changing it. We had initialized all out hyper-parameters with the same values as specified by [2] in their code. Fine-tuning the hyper-parameters could give small gains in performance but due to time constraints we could not tune all hyper-parameters (One cycle of training on any model takes about 8 hours).

Models	Accuracy (%)
LSTMIMG [2] (Baseline)	**54.06**
NLOR_C	51.59
NLOR	52.25
NLOR_MC	52.94
NLOR_3B	53.62
NLOR_M	53.97
Sal	52.63
NLOR_3B_Sal	52.74
Sal_M	53.96
NLOR_Sal	**54.51**

Table 2: Performance of our models in comparison to baseline approach on the MSCOCO-VQA dataset [2]

5.2.2 Performance over Question Types

We analyze the performance on 65 different question types of our various models in comparison to the baseline [2]. The models which only use NLOR outperform in combina-



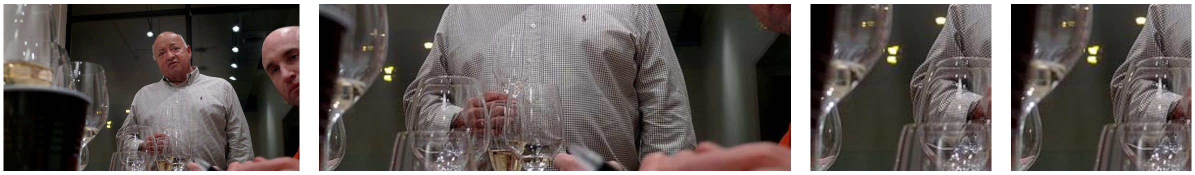
(a) Original Image

(b) What shape is the bench seat?

(c) Is there a shadow?

(d) Is this one bench or multiple benches?

Figure 12: An example of the top-1 bounding box obtained from the Natural Language Object Retrieval (NLOR) Pipeline for a sample image (a) pertaining to different questions based on it (b), (c), (d). It is interesting to note that in (b), the bounding box rated with the highest score zooms in on the shape of the seat of the bench, latching on to a unique structure on the bench, though not the bench as a whole from a distant view; in (c) the selected bounding box captures a region of the image where the shadows are very obvious, such as close to the leg of the bench and on the seat of the bench. There are shadows of trees on the road as well, but interestingly the NLOR pipeline extracts a region which has very prominent shadows; in (d), the bounding box manages to include all the benches such that the related question could be answered correctly. The extracted bounding boxes seem appropriate for most images (and at the least not completely incorrect), whether they will improve performance on the VQA task remains to be seem.



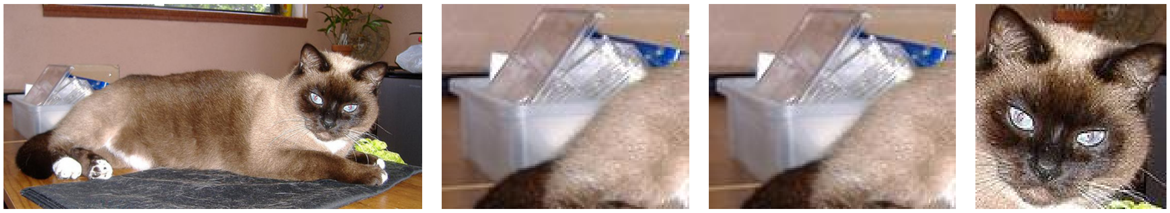
(a) Original Image

(b) What are the men doing?

(c) What kind of glasses are on the table?

(d) How old are these the men?

Figure 13: Another example of the top-1 bounding box obtained from the Natural Language Object Retrieval (NLOR) Pipeline for a sample image (a) pertaining to different questions based on it (b), (c), (d). Interestingly for (b), the retrieved bounding box focuses on the men's hands holding the glass suggestive of the act of drinking; for (c) the retrieved bounding box captures an area mostly occupied by transparent glasses, which is impressive, though it is arguable if it is the most distinctive bounding box capturing a nice view of the shape the glass; for (d) the retrieved bounding box does not look appropriate at all since probably a subregion capturing the faces of the men would be more indicative of their age versus just a hand of one man and primarily the glasses.



(a) Original Image

(b) What breed of cat is in the photo?

(c) What is the cat on?

(d) What color is the cat's eyes?

Figure 14: One more example of the top-1 bounding box obtained from the Natural Language Object Retrieval (NLOR) Pipeline for a sample image (a) pertaining to different questions based on it (b), (c), (d). For questions (b) and (c), the retrieved bounding boxes seem inappropriate and the object in focus in these retrieved regions is just part of the background. The questions are more relevant to the cat which is not captured here; for question (d), the bounding box focuses on the face of the cat and the eyes are well highlighted, so the extracted region seems appropriate.

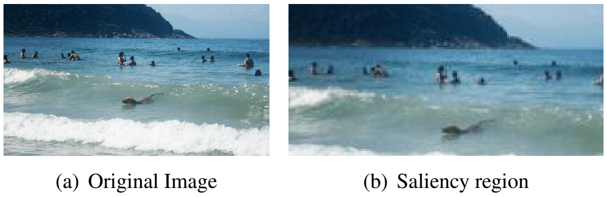

(a) Original Image (b) Saliency region

Figure 15: In this example, most of the image, except the surf in the front is captured in the salient region. Questions with this image -
– What animal is in the water?
– What is the animal in the water?
– How many people are present?
The first two can be answered from the region given by saliency. It is possible that some of the people in the background have been cropped out which would make the third question difficult to answer correctly.

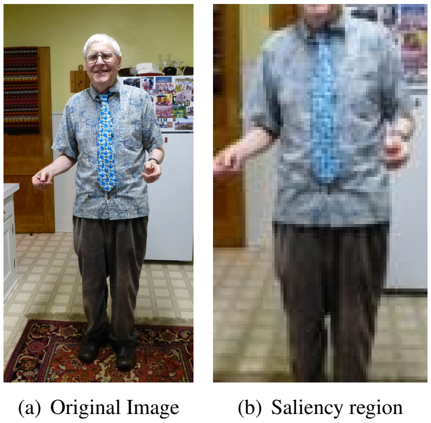

(a) Original Image (b) Saliency region

Figure 16: In this example, saliency seems to focus on the shirt and tie, possibly because of the interesting pattern. Questions with this image -
– What room is this?
– What color is the tie?
– Is this the proper shirt to wear with a tie?
This region would be suitable for the second and third questions but the first would probably be better answered by observing background regions such as the fridge in the original image.

tion on roughly 65% of the questions compared to the baseline (Fig. 28). This chart shows that the models are learning features which capture complementary information and are not performing well on questions which are subsumed completely by other models. We also compare the best performing model which uses only NLOR outputs and no saliency information (NLOR_M) with the baseline in a similar manner (Fig. 30). The baseline is overall better than

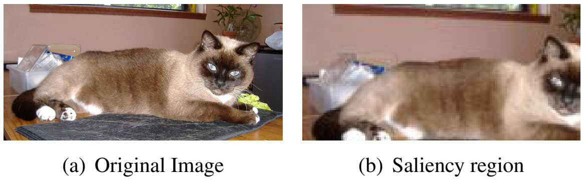

(a) Original Image (b) Saliency region

Figure 17: In this example, most of the body of the cat is captured. It is interesting that the entire face is not captured. Questions with this image -
– What breed of cat is in the photo?
– What is the cat on?
– What color is the cat's eyes?
Since the face is not fully captured, and the region capturing the rug the cat is lying on is completely cropped out of the saliency region, this set of questions will probably be difficult to answer given the saliency region.

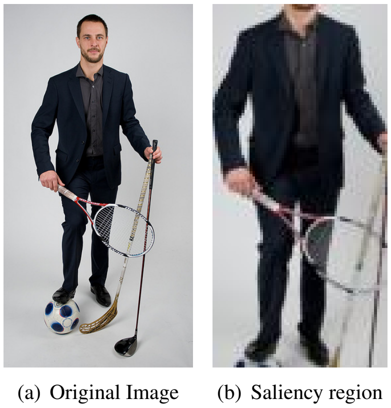
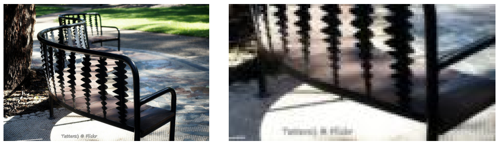

(a) Original Image (b) Saliency region

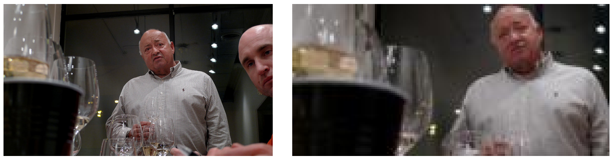
Figure 18: The region captured by saliency here is a little odd in that it neither captures the head of the man, nor the tips of the sports equipment at his feet. This is probably because the colour contrast between his suit and the background probably outweighs other features. Questions with this image -
– How many different types of sports equipment is he holding?
– Would someone typically use all these together?
– Can he fly on his own?
The first two questions require all the sports equipment to be identified. This would probably be difficult from the saliency region. However, what makes the second and third questions more difficult for a VQA system is that common sense is required to answer them.

this model and this is reflected in the number of question types the baseline outperforms on. The baseline does better than Model NLOR_M on 58.5% of the question types and Model NLOR_M performs better on 41.5%. In general we



(a) Original Image (b) Saliency region

Figure 19: Here, saliency captures the main part of the larger bench in the image. Questions with this image -
– What shape is the bench seat?
– Is there a shadow?
– Is this one bench or multiple benches?
While the saliency region would be able to answer the first two questions, it would not be able to answer the third.



(a) Original Image (b) Saliency region

Figure 20: The saliency region in this image consists of the more prominent man as well as some of the wine glasses. Questions with this image -
– What are the men doing?
– What kind of glasses are on the table?
– How old are these men?
It should be possible for the system to answer these questions from the saliency region, although the third question might require some commonsense knowledge.

find that NLOR related models do better on questions which include a mention to objects in the image within the question such as 'what is the woman', 'is there a', 'does that', 'is this person' etc. versus questions like 'where is the', 'how many', 'how many people are' on which the baseline does better. This intuitively goes well with what the NLOR pipeline is originally trained on - queries describing objects in the image.

We also do a similar analysis on the models which use salient regions of the image as input. In Fig **??**, we compute on what percentage of question types the various models outperform compared to the others. These models in general do better than the ones using only saliency. Combined they outperform the baseline on 60% of the question types. Here also we see that these models capture complementary information compared to each other and the baseline. In Fig. **??**, we find that our best performing model (Model NLOR_Sal) does well on 50.8% of the question types compared to the baseline.

5.2.3 Performance over Answer Types

A comparison of performance of various types can be seen in table 3. Among models that do not use saliency, the *NLOR_M* model performs the best, which is a similar trend to the overall performance. However, among all models, the model using only saliency performs the best on number type questions. The model using saliency, NLOR and the original image - model *NLOR_Sal* still outperforms the baseline on all answer types.

Models	Yes/No (%)	Number (%)	Other (%)
LSTMIMG [2] (Baseline)	79.78	32.99	40.29
NLOR_C	78.6	32.32	36.21
NLOR	79.44	32.51	36.86
NLOR_MC	79.54	32.4	38.18
NLOR_3B	79.57	31.75	39.68
NLOR_M	**79.79**	**32.94**	**39.92**
Sal	79.45	32.96	37.5
NLOR_3B_Sal	79.01	31.37	38.45
Sal_M	79.64	**33.32**	39.92
NLOR_Sal	**79.81**	33.19	**40.93**

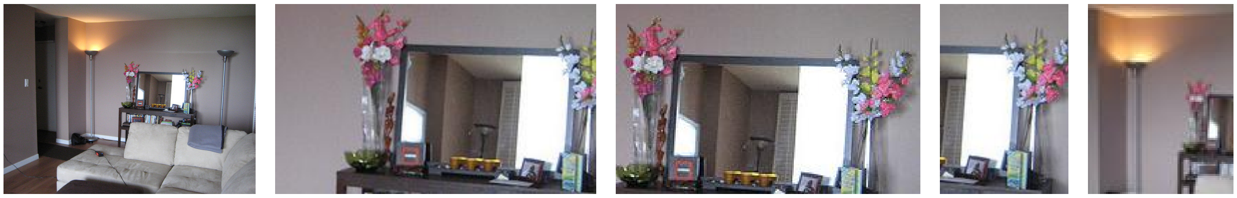
Table 3: Performance per answer type of our models in comparison to baseline and state of the art approaches on the MSCOCO-VQA dataset [2]

6. Discussion and Conclusions

We find that NLOR is not particularly helpful in improving performance on VQA overall, although a comparison over performance on individual question types indicates that NLOR does capture some information complementary to the baseline. For many image-question pairs, the regions returned by NLOR do not appear to be qualitatively useful, which also limits the performance of NLOR based models. Saliency also does not seem to perform well by itself or with the original image. The regions when examined qualitatively sometimes seem to capture information relevant to questions but failure cases are also easy to find.

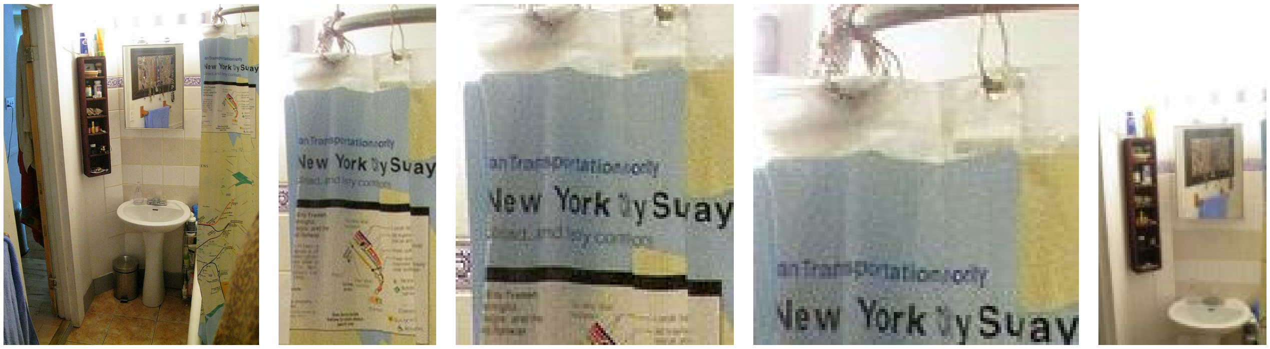
However, the combination of information form both NLOR and saliency allows us to improve over the baseline. This suggests that these two methods capture complementary information that is often useful to answer questions. This model possibly performs better as it is able to exploit cases where either NLOR or saliency identifies a relevant region.

We also find that using multiple regions from NLOR is not very helpful. One possible reason for this could be that



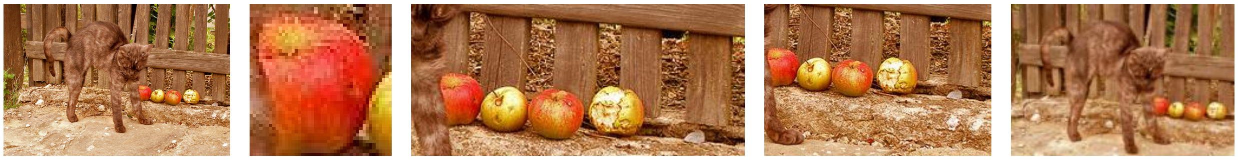
(a) Original Image (b) NLOR region 1 (c) NLOR region2 (d) NLOR region 3 (e) Saliency region

Figure 21: **Question**: What are the colorful objects to either side of the mirror? **Ground truth**: flowers
Baseline: light, **NLOR_M**: light, **NLOR_3B**: fan, **Sal_M**: tv, **NLOR_Sal**: flowers
In this example, both NLOR and saliency select regions that contain the answer to the question, although tighter regions are possible. We find that some models answer light, which could be due to the reflected light from the mirror. However, model using both NLOR and saliency gets the correct answer.



(a) Original Image (b) NLOR region 1 (c) NLOR region2 (d) NLOR region 3 (e) Saliency region

Figure 22: **Question**: What subway is on the curtain? **Ground truth**: new york
Baseline: obama, **NLOR_M**: dunkin donuts, **NLOR_3B**: coca cola, **Sal_M**: none, **NLOR_Sal**: new york
In this case, NLOR focusses on the correct region in all three bounding boxes but only the model in combination from saliency gets this answer correct, despite the saliency region not being relevant to the answer.

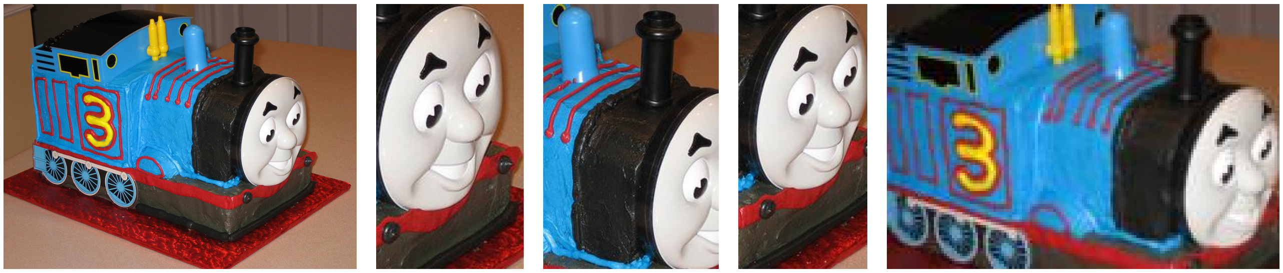


(a) Original Image (b) NLOR region 1 (c) NLOR region2 (d) NLOR region 3 (e) Saliency region

Figure 23: **Question**: What type of fruit is in the photo? **Ground truth**: apple
Baseline: elephant, **NLOR_M**: elephant, **NLOR_3B**: elephant, **Sal_M**: apple, **NLOR_Sal**: bananas
In this example, NLOR identifies good regions, with all of the three regions focussing on one or more apples. However, possibly because the texture is unusual, only the model with both NLOR and saliency picks up on this. Saliency itself seems to crop out only a very small fraction of the image.

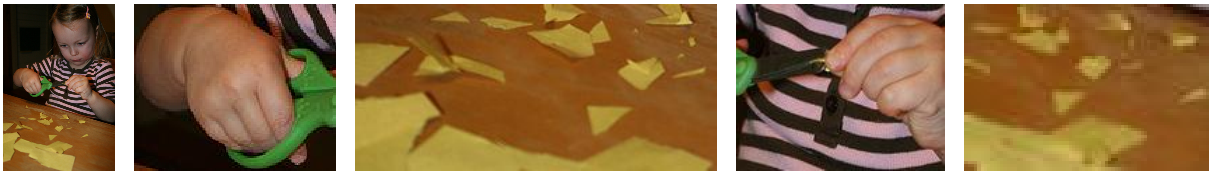
the increased number of parameters in this model results in overfitting.

The team which released the VQA data [2] has recently also released a plot confirming that their model is data starved. VQA accuracy vs. Log(training data samples) is a straight line as can be seen in Fig. 32, and hence the parameter estimates are possibly not good enough. This could be one reason why the performance does not improve



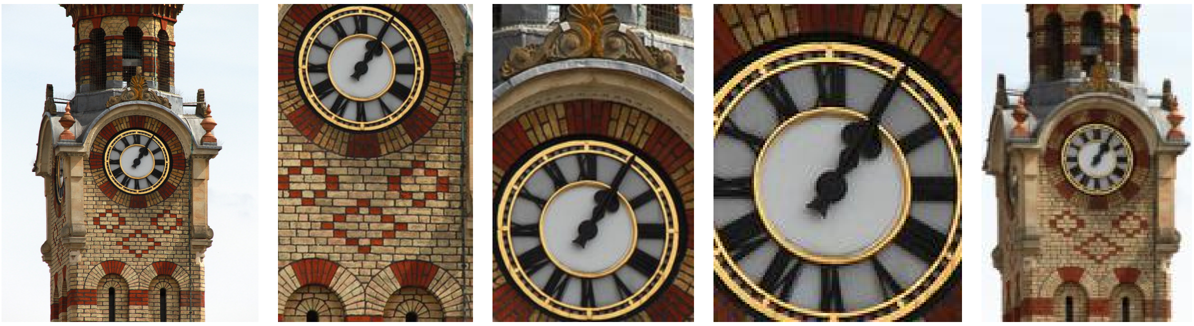
(a) Original Image (b) NLOR region 1 (c) NLOR region2 (d) NLOR region 3 (e) Saliency region

Figure 24: **Question**: What number is on the cake? **Ground truth**: 3
Baseline: 2, **NLOR_M**: 1, **NLOR_3B**: 2, **Sal_M**: 3, **NLOR_Sal**: 4
In this example, NLOR identifies very poor regions, possibly because the train shape makes it less obvious what is the cake in the image. However, the region from saliency contains the answer to the question. We find that only the model Sal_M which does not use either the original image or NLOR regions gets the correct answer.



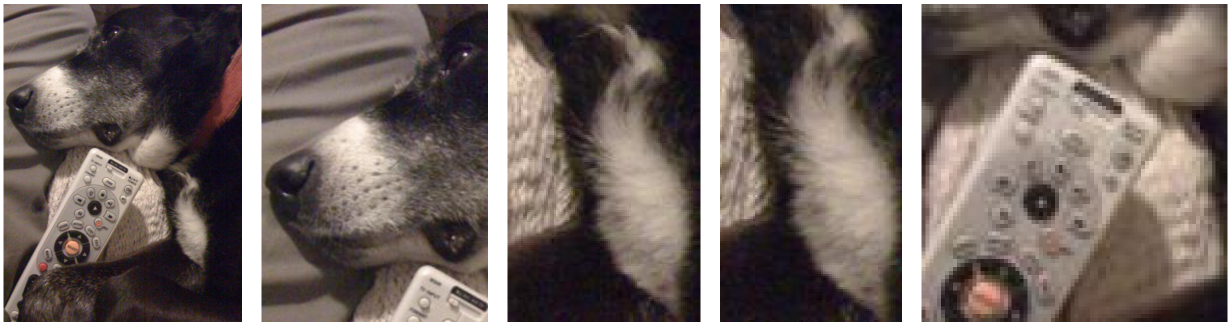
(a) Original Image (b) NLOR region 1 (c) NLOR region2 (d) NLOR region 3 (e) Saliency region

Figure 25: **Question**: What color handle do the scissors have? **Ground truth**: green
Baseline: red, **NLOR_M**: blue, **NLOR_3B**: green, **Sal_M**: blue, **NLOR_Sal**: pink
In this example, the top NLOR region focusses very well on the handle of the scissors. The region from saliency however, does not contain the answer. We find that only the model using all three NLOR regions gets the answer correct, which is unusual since the second and third bounding box from NLOR don't seem to be very useful.



(a) Original Image (b) NLOR region 1 (c) NLOR region2 (d) NLOR region 3 (e) Saliency region

Figure 26: **Question**: What color hands does the clock have? **Ground truth**: black
Baseline: gold, **NLOR_M**: black, **NLOR_3B**: gold, **Sal_M**: gold, **NLOR_Sal**: gold
In this example, all the NLOR regions seem to focus on the clock face, however, instead of the model that uses the three regions, the one using only a single region is the only model to get the answer correct. All the other models guess gold, possibly because there is some gold colour on the clock face, and possibly also because a lot of wristwatches have gold hands.



(a) Original Image (b) NLOR region 1 (c) NLOR region2 (d) NLOR region 3 (e) Saliency region

Figure 27: **Question**: What color collar does the dog have? **Ground truth**: red
Baseline: red, **NLOR_M**: blue, **NLOR_3B**: white, **Sal_M**: blue, **NLOR_Sal**: pink
In this example, neither NLOR nor saliency identify the correct region. The baseline model is the only one which gets it correct.

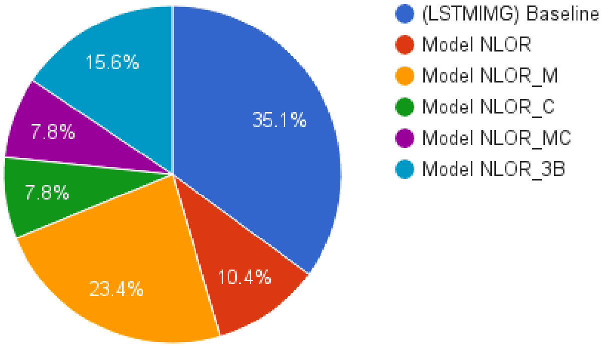


Figure 28: A distribution of question types our proposed NLOR related models and the baseline respectively outperform all others on

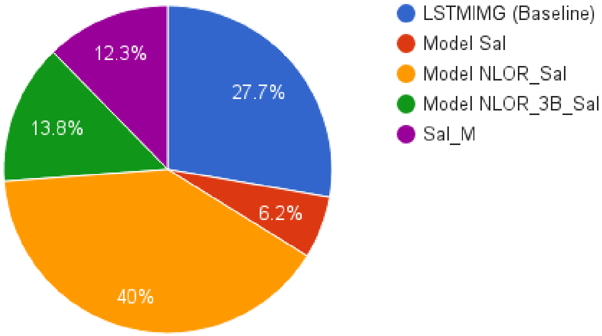


Figure 29: A distribution of question types our proposed saliency related models and the baseline respectively outperform all others on

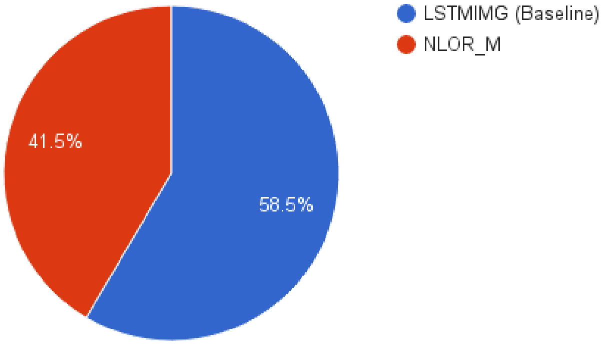


Figure 30: A distribution of question types our best performing NLOR related model and the baseline respectively outperform each other on

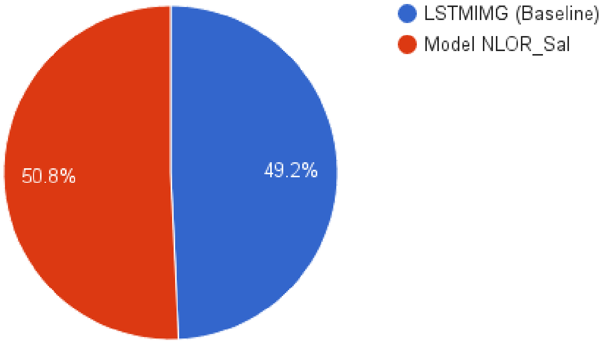


Figure 31: A distribution of question types our best performing saliency related model and the baseline respectively outperform each other on

since there are not enough image-question pairs to generalize from. In our analysis, we observe the training error for the model using NLOR and saliency regions is lower compared to the baseline, yet we do not see significant improvement on the validation data. One intuition behind this could be the bias present in the dataset towards questions which

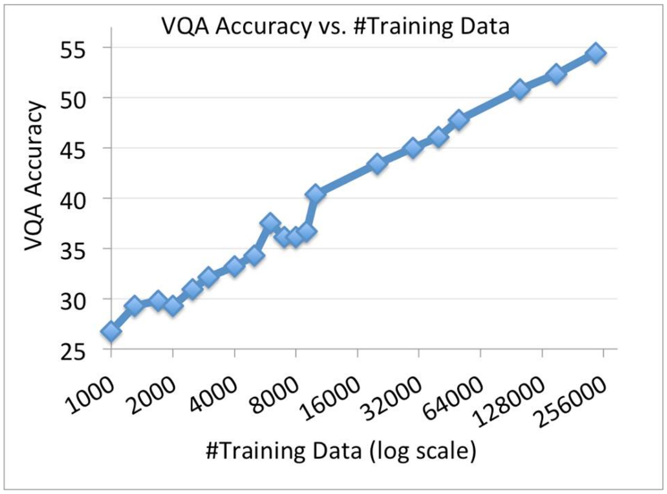


Figure 32: VQA accuracy vs. Log(Number of Training Data samples)

have common answers such as yes/no and numbers. Most of such questions can be answered correctly using the trivial approach of selecting the most common answer. Hence the number of image-question pairs from which the model should generalize significantly reduces and the model probably overfits to these common cases.

7. Future Work

There are many possible directions to extend our proposed approach. One possibility is that the Natural Language Object Retrieval pipeline [6] could be retrained (instead of the pretrained model made available by [6]) with a dataset such as Visual7W [25] which has bounding box annotations corresponding to regions containing the answers to image related questions. Another option is to extract noun/verb phrases from the question and use those as the input for the NLOR model. Since NLOR is trained on queries describing an object in the image, using questions instead of natural language descriptions at test time may not be lead to exploiting the full potential of their model. Tuning of other hyperparameters beyond the learning rate may also improve performance. Another possibility would be to use NLOR as an optional module, which is activated only for the question types on which NLOR is seen to benefit. On the saliency side, using a neural model of saliency such as [23] that exploits higher level cues instead of low level cues such as colour and contrast might better capture regions of the image that humans are likely to ask questions about.

References

[1] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Deep compositional question answering with neural module networks. *arXiv preprint arXiv:1511.02799*, 2015.

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *International Conference on Computer Vision (ICCV)*, 2015.

[3] K. Chen, J. Wang, L.-C. Chen, H. Gao, W. Xu, and R. Nevatia. Abc-cnn: An attention based convolutional neural network for visual question answering. *arXiv preprint arXiv:1511.05960*, 2015.

[4] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. Are you talking to a machine? dataset and methods for multilingual image question answering. *arXiv preprint arXiv:1505.05612*, 2015.

[5] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, Nov. 1997.

[6] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell. Natural language object retrieval. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[7] I. Ilievski, S. Yan, and J. Feng. A focused dynamic attention model for visual question answering. arXiv:1604.01485.

[8] A. Jiang, F. Wang, F. Porikli, and Y. Li. Compositional memory for visual question answering. *arXiv preprint arXiv:1511.05676*, 2015.

[9] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H. Y. Shum. Learning to detect a salient object. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(2):353–367, Feb 2011.

[10] L. Ma, Z. Lu, and H. Li. Learning to answer questions from image using convolutional neural network. *arXiv preprint arXiv:1506.00333*, 2015.

[11] M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in Neural Information Processing Systems*, pages 1682–1690, 2014.

[12] M. Malinowski and M. Fritz. Towards a visual turing challenge. *arXiv preprint arXiv:1410.8027*, 2014.

[13] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–9, 2015.

[14] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119, 2013.

[15] H. Noh, P. H. Seo, and B. Han. Image question answering using convolutional neural network with dynamic parameter prediction. *arXiv preprint arXiv:1511.05756*, 2015.

[16] M. Ren, R. Kiros, and R. Zemel. Exploring models and data for image question answering. In *Advances in Neural Information Processing Systems*, pages 2935–2943, 2015.

[17] K. J. Shih, S. Singh, and D. Hoiem. Where to look: Focus regions for visual question answering. *arXiv preprint arXiv:1511.07394*, 2015.

[18] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[19] Q. Wu, P. Wang, C. Shen, A. v. d. Hengel, and A. Dick. Ask me anything: Free-form visual question answering based on knowledge from external sources. *arXiv preprint arXiv:1511.06973*, 2015.

[20] H. Xu and K. Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. *arXiv preprint arXiv:1511.05234*, 2015.

[21] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. *arXiv preprint arXiv:1511.02274*, 2015.

[22] L. Yu, E. Park, A. C. Berg, and T. L. Berg. Visual madlibs: Fill in the blank image generation and question answering. *arXiv preprint arXiv:1506.00278*, 2015.

[23] J. Zhang, S. Ma, M. Sameki, S. Sclaroff, M. Betke, Z. Lin, X. Shen, B. Price, and R. Mech. Salient object subitizing. In *CVPR*, 2015.

[24] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus. Simple baseline for visual question answering. *arXiv preprint arXiv:1512.02167*, 2015.

[25] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7w: Grounded question answering in images. *arXiv preprint arXiv:1511.03416*, 2015.

[26] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *Computer Vision–ECCV 2014*, pages 391–405. Springer, 2014.