APPLYING VCR MODEL ON FVQA DATA

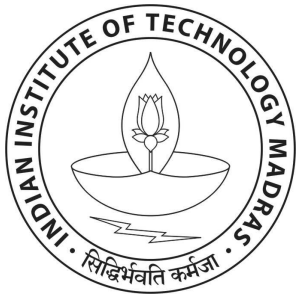
A Project Report

submitted by

SANKALP SAOJI

in partial fulfilment of the requirements
for the award of the degree of

BACHELOR OF TECHNOLOGY



DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY MADRAS.

JUNE 2020

REPORT CERTIFICATE

This is to certify that the report titled **APPLYING VCR MODEL ON FVQA DATA**, submitted by **Sankalp Saoji**, to the Indian Institute of Technology, Madras, for the award of the degree of **Bachelor of Technology**, is a bona fide record of the research work done by him under our supervision. The contents of this report, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

B.Tech Project Guide
Dr. Anurag Mittal
Department of Computer Science
IIT-Madras, 600 036

Place: Chennai

Date: 25th June 2020

ABSTRACT

In my project on Visual Question Answering, I have referred to the papers - FVQA (**FVQA: Fact-based Visual Question Answering by Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick and Anton van den Hengel**) and VCR (**From Recognition to Cognition: Visual Commonsense Reasoning by Rowan Zellers, Yonatan Bisk, Ali Farhadi and Yejin Choi**). Both the VCR and FVQA papers are concerned with reasoning along with VQA and both of these papers have models which take into account some fact or reason for the answer predicted to the question asked about the image. So, I have tried to implement a model based on VCR on the FVQA dataset. VCR model was used to identify the reason for an answer amongst the multiple options given to it. I have tried to select certain facts in the same way from the FVQA dataset which support the answer. But, due to lack of computing power, I have trained and tested on a limited amount of data. I have got an accuracy of about **52.121%** on the test set.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Visual Question Answering has got a lot of interest these days from both the natural language processing and computer vision communities. That is because it associates two different pieces of information together. The field has made significant progress on recognition-level building blocks. The set of questions that a VQA method is able to answer are one of its key features, and limitations. Asking a method a question that is outside its scope will lead to a failure to answer, or worse, to a random answer. Much of the existing VQA effort has been focused on questions which can be answered by the direct analysis of the question and image, on the basis of a large training set. Most of the models and datasets that we have now, focus on questions which can be answered by solely analysing the question and image involved without giving any regard to anything else. The datasets exclude questions whose answers need common sense as possessed by human beings or any factual knowledge that can be known by the model. Below mentioned are the two different approaches to the problem of VQA which have datasets with one more factor of reasoning along with the image, question and answer triplets.

1.1 FVQA (Fact-based VQA)

FVQA is a dataset which contains questions that require external information to answer. Unlike a conventional visual question answering dataset, which contains image-question-answer triplets, in FVQA dataset we have additional image-question-answer-supporting fact tuples. It thus promotes deeper reasoning. For example, given an image with a cat and a dog, and the question 'Which animal in the image is able to climb trees?', the answer is 'cat'. The required supporting-fact for answering this question is <cat is capable of climbing trees>. By providing supporting facts, the dataset supports answering complex questions, even if all of the information required to answer the question is not depicted in the image. Moreover, it supports explicit reasoning in visual question answering because it gives us an indication as to how a method might derive

Figure 1.1: An example visual-based question from the FVQA dataset that requires both visual and common-sense knowledge to answer

an answer to a particular question. This information can be used in answer inference, to search for other appropriate facts, or to evaluate answers which include an inference chain.

1.2 VCR (Visual Commonsense Reasoning)

When a human being looks at an image, he not only sees the objects in it but forms deeper connections as he has some prior information with the objects involved in there. This is very difficult for today's systems as it requires common knowledge about the world and cognition. This task is dealt under Visual Commonsense Reasoning. A new dataset, VCR consisting of 290k multiple choice QA problems derived from 110k movie scenes is used. When a question about an image is given to the model, the model must answer the question in a correct way and then provide a rationale for its answer. The machine must answer a question that requires a thorough understanding of the visual world evoked by the image. Moreover, the machine must provide a rationale justifying why that answer is true, referring to the details of the scene, as well as background knowledge about how the world works. These questions, answers, and rationales are expressed using a mixture of rich natural language as well as explicit references to image regions. Given a question along with four answer choices, a model

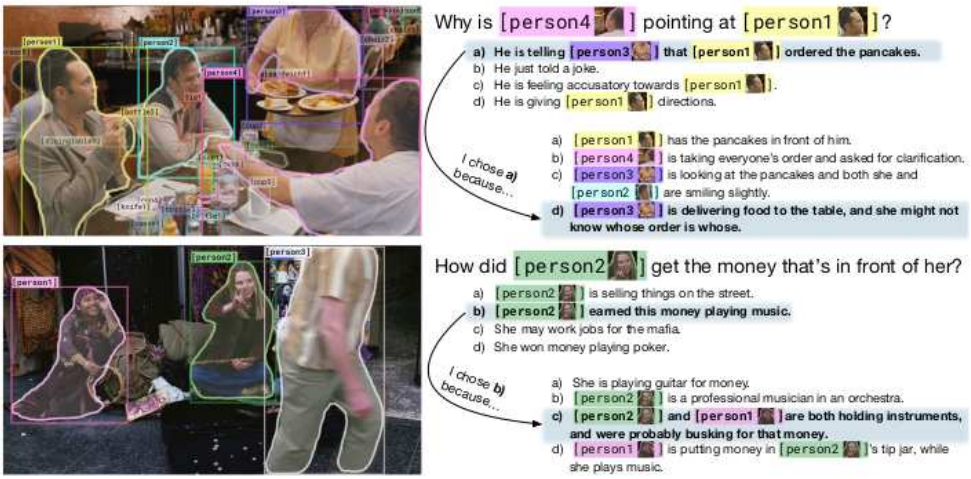


Figure 1: VCR: Given an image, a list of regions, and a question, a model must answer the question and provide a *rationale* explaining why its answer is right. Our questions challenge computer vision systems to go beyond recognition-level understanding, towards a higher-order cognitive and commonsense understanding of the world depicted by the image.

Figure 1.2: Given an image, a list of regions, and a question, a model must answer the question and provide a rationale explaining why its answer is right

must first select the right answer. If its answer was correct, then it is provided four rationale choices (that could purportedly justify its correct answer), and it must select the correct rationale. We call this QAR as for the model prediction to be correct requires both the chosen answer and then the chosen rationale to be correct. Our task can be decomposed into two multiple-choice sub-tasks, that correspond to answering (QA) and justification (QAR) respectively.

1.3 Dataset

The data used for this task was the FVQA dataset. The FVQA dataset typically looks as shown in Figure 1.3.

Any single data point from the dataset contains information as below:

1] fact_surface = 'You are likely to find [[a trumpet]] in [[a jazz club]]' (fact_surface: the fact associated with the answer)

2] ans_source = 'image' (ans_source: the source of the answer)

3] answer = 'trumpet' (answer: answer of the question)

4] question = 'Which object can be found in a jazz club' (question: question asso-

Dataset	Number of images	Number of questions	Num. question categories	Average quest. length	Average ans. length	Knowledge Bases	Supporting-Facts
DAQUAR [12]	1,449	12,468	4	11.5	1.2	-	-
COCO-QA [9]	117,684	117,684	4	8.6	1.0	-	-
VQA-real [5]	204,721	614,163	20+	6.2	1.1	-	-
Visual Genome [11]	108,000	1,445,322	7	5.7	1.8	-	-
Visual7W [10]	47,300	327,939	7	6.9	1.1	-	-
Visual Madlibs [8]	10,738	360,001	12	6.9	2.0	-	-
VQA-abstract [5]	50,000	150,000	20+	6.2	1.1	-	-
VQA-balanced [67]	15,623	33,379	1	6.2	1.0	-	-
KB-VQA [61]	700	2,402	23	6.8	2.0	1	-
Ours (FVQA)	2,190	5,826	32	9.5	1.2	3	✓

Figure 1.3: Major datasets for VQA and their main characteristics


```
{"270": {"fact_surface": "You are likely to find [[a trumpet]] in [[a jazz club]]",
"ans_source": "image", "answer": "trumpet", "question": "Which object can be found
in a jazz club", "img_file": "ILSVRC2012_test_00050748.JPEG", "visual_concept":
"obj", "kb_source": "conceptnet", "fact":
["conceptnet/e/f768f157e4446dd594536f8ef02681515586ba2d"], "question_id": "270"},
"271": {"fact_surface": "[[lipstick]] belongs to the category of [[Cosmetics]]",
"ans_source": "image", "answer": "lipstick", "question": "Tell me the name of the
cosmetics shown in this image?", "img_file": "ILSVRC2012_test_00000444.JPEG",
"visual_concept": "obj", "kb_source": "dbpedia", "fact": ["dbpedia/8657"],
"question_id": "271"}, "272": {"fact_surface": "[[Lipstick]] is for [[coloring the
lips]]", "ans_source": "kb", "answer": "coloring the lips", "question": "What is
the object shown in this image used for", "img_file":
"ILSVRC2012_test_00000444.JPEG", "visual_concept": "obj", "kb_source":
"conceptnet", "fact": ["conceptnet/e/18b413dee0ebe3cadef337f364baa61649f36b2a"],
"question_id": "272"}, "273": {"fact_surface": "You are likely to find [[lipstick]]
in [[a makeup cabinet]]", "ans_source": "kb", "answer": "a makeup cabinet",
"question": "Where can you find the object in this image", "img_file":
"ILSVRC2012_test_00000444.JPEG", "visual_concept": "obj", "kb_source":
"conceptnet", "fact": ["conceptnet/e/c8399f11b843cb0fa84dbfe865954212660a801c"],
"question_id": "273"}, "274": {"fact_surface": "[[A kite]] has [[a tail]]",
"ans_source": "image", "answer": "kite", "question": "Which object in this image
has a tail", "img_file": "COCO_val2014_000000005599.jpg", "visual_concept": "obj",
"kb_source": "conceptnet", "fact":
["conceptnet/e/1e9f66df12d446019dc02db7fe054b87c114e8d3"], "question_id": "274"},
"275": {"fact_surface": "A [[sandwich]] is a [[meal commonly eaten for lunch]]",
"ans_source": "image", "answer": "sandwich", "question": "what object in this image
is commonly eaten for lunch?", "img_file": "COCO_val2014_000000015079.jpg",
"visual_concept": "obj", "kb_source": "conceptnet", "fact":
["conceptnet/e/03ae54517f861f4d9130e4cfce1b4f23ad3363c8"], "question_id": "275"},
```


Figure 1.4: FVQA Dataset

Which animal in this image has stripes?

(HasA, Object, Image)
'stripes'

zebras have stripes

Which transportation way in this image is cheaper than taxi?

(Cheaper, Object, Image)
'taxi'

bus are cheaper than taxi

Figure 1.5: Dataset Example 1

Which furniture in this image can I lie on?	What animal in this image are pulling carriage?
(UsedFor,Object,Image) 'lie on'	(CapableOf,Object,Image) 'pulling carriage'
a sofa is usually to sit or lie on	horses sometimes pull carriages

Figure 1.6: Dataset Example 2

ciated with the image)

5] img_file = 'ILSVRC2012_test_00050748.JPEG' (img_file: the image which was used and on whom the question was asked)

6] visual_concept = 'obj' (visual_concept: tells what the visual concept was)

7] kb_source = 'conceptnet' (kb_source: the source of the fact used)

8] fact = 'conceptnet/e/f768f157e4446dd594536f8ef02681515586ba2d' (fact: the fact which was extracted)

9] question ID = '270' (question ID: the index of the question - because same image can have many questions associated with it)

CHAPTER 2

APPROACH

The approach of Figure 2.1 was the one used in the original paper. As shown from the figure, we can see that the process takes place as per the following steps:

1) First, the image is passed through object detector, scene classifier and then through an attribute classifier. After going through these, we get the extracted visual concepts. Object Detector gives out the objects present in the image. Scene Classifier gives out the details about the surrounding in which the objects are located. Then, lastly, attribute classifier gives out information about the action taking place.

2) Now, the question asked is broken down through question query mapping by passing through an LSTM. The KB query is thus obtained.

3) Lastly, the KB query and the object, scene and action details are all searched about in the Knowledge Base which in the case of the paper was a combination of DBPedia, ConceptNet and WebChild. Thus, we arrive at our answer.

But, this was the approach of the FVQA paper. I have implemented the VCR approach as mentioned below on the FVQA dataset.

Recognition to Cognition Network (R2C), a new model for visual commonsense reasoning is used. The model is given an image, a set of objects 'o', a query 'q', and a set of responses 'r' of which exactly one is correct. This was as per the VCR paper. I have given the set of facts from FVQA dataset as responses to the VCR model. All the facts are given to the model to choose from. The facts matter help in obtaining the correct answers unlike the VCR case where answers were selected first and then the reasons. The query q and fact choices are all expressed in terms of a mixture of natural language and pointing to image regions. For our image features, we use ResNet50. To obtain strong representations for language, we used BERT representations. BERT is applied over the entire question and answers, and we extract a feature vector from the second-to- last layer for each word. There are three main steps in the R2C model which are further explained below:

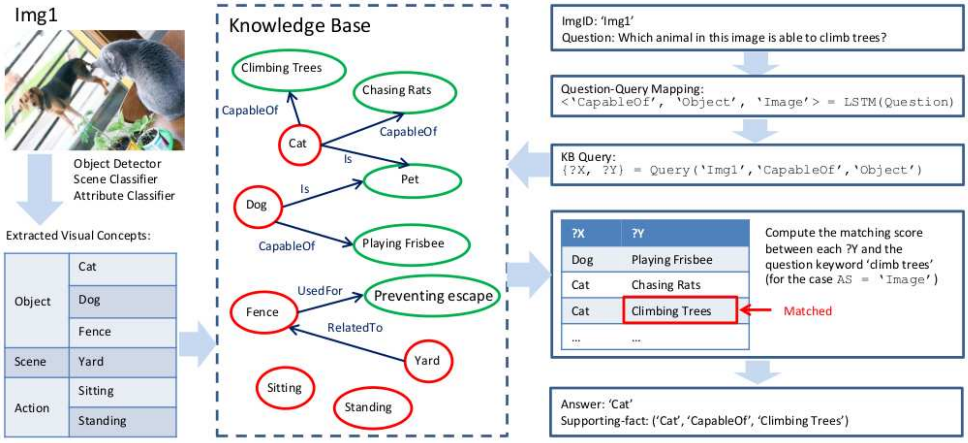


Figure 2.1: The reasoning process of the paper's VQA approach. The visual concepts (objects, scene, attributes) of the input image are extracted using trained models, which are further linked to the corresponding semantic entities

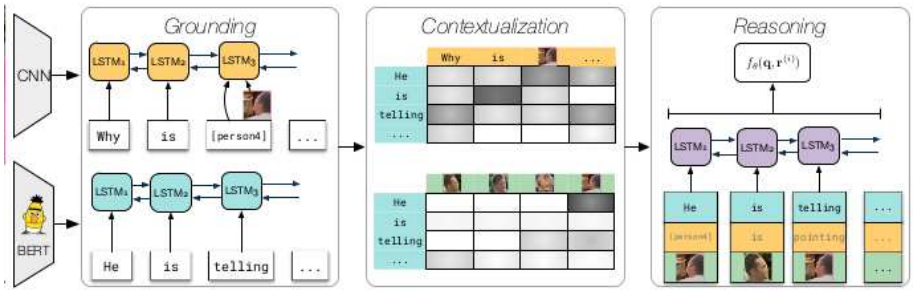


Figure 2.2: In R2C model, we break the challenge of Visual Commonsense Reasoning into three components: grounding the query and response, contextualizing the response within the context of the query and the entire image, and performing additional reasoning steps on top of this rich representation

STEPS:

1) Grounding

The grounding module learns a joint image-language representation for each token in a sequence. At the core of the grounding module is a bidirectional LSTM. We then use a CNN to learn object-level features.

2) Contextualization

Given a grounded representation of the query and response, we use attention mechanisms to contextualize these sentences with respect to each other and the image context. To contextualize an answer with the image, including implicitly relevant objects that have not been picked up from the grounding stage, we perform another bilinear attention between the response r and each object's image features.

3) Reasoning

Lastly, we allow the model to reason over the response, attended query and objects. We accomplish this using a bidirectional LSTM that is given as context qi, ri, and oi for each position i. For better gradient flow through the network, we concatenate the output of the reasoning LSTM along with the question and answer representations for each timestep. The resulting sequence is max-pooled and passed through a multilayer perceptron, which predicts for the query-response compatibility.

CHAPTER 3

RESULTS

On the FVQA dataset, human subjects were only allowed to provide one answer to one question, so there is no Top-3 and Top-10 evaluations for the human performance. Three variants were implemented in the paper. 'gt-QQmaping' uses the ground truth question-query mapping, while 'top-1-QQmaping' and 'top-3-QQmaping' use the top-1 and top-3 predicted question-query mapping. This all data is shown in Figure 3.1.

Due to limitation in computing power, I have just used about 1100 sets of images, questions and facts. On these, I have trained on about 70% data points while used the rest of the data for testing. I have tested on 330 images out of which 172 showed correct output. By correct output, I mean the final output and not the fact.

Accuracy = (172/330)*100 = **52.121%**

Accuracies	**Values**
Human Accuracy	77.99%
Paper's Accuracy	63.63%
My model's Accuracy	52.121%

Method	Overall Acc. ± Std (%)		
	Top-1	**Top-3**	**Top-10**
SVM-Question	10.37 ± 0.80	20.72 ± 0.58	34.63 ± 1.19
SVM-Image	18.41 ± 1.07	32.42 ± 1.06	47.53 ± 1.02
SVM-Question+Image	18.89 ± 0.91	32.78 ± 0.90	48.13 ± 0.73
LSTM-Question	10.45 ± 0.57	19.02 ± 0.74	31.64 ± 0.93
LSTM-Image	20.55 ± 0.81	36.01 ± 1.45	55.74 ± 2.28
LSTM-Question+Image	22.97 ± 0.64	36.76 ± 1.22	54.19 ± 2.45
LSTM-Question+Image+Pre-VQA	24.98 ± 0.60	40.40 ± 1.05	57.27 ± 1.29
Hie-Question+Image	33.70 ± 1.18	50.00 ± 0.78	64.08 ± 0.57
Hie-Question+Image+Pre-VQA	43.14 ± 0.61	59.44 ± 0.34	$\mathbf{72.20 \pm 0.39}$
Ours, gt-QQmaping[‡]	63.63 ± 0.73	71.30 ± 0.78	72.55 ± 0.79
Ours, top-1-QQmaping	52.56 ± 1.03	59.72 ± 0.82	60.58 ± 0.86
Ours, top-3-QQmaping	$\mathbf{56.91 \pm 0.99}$	$\mathbf{64.65 \pm 1.05}$	65.54 ± 1.06
Ensemble	58.76 ± 0.92	-	-
Human	77.99 ± 0.75	-	-

Figure 3.1: Accuracies on FVQA dataset

REFERENCES

1. FVQA: Fact-based Visual Question Answering Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, Anthony Dick

2. From Recognition to Cognition: Visual Commonsense Reasoning Rowan Zellers, Yonatan Bisk, Ali Farhadi, Yejin Choi