Knowledge-based Visual Question Answering

Lucky, Rashi Verma

Abstract

Visual Question Answering (VQA) is an important task in the field of Computer Vision and Natural Language Processing. The task of VQA is basically producing answer in natural language to a question asked in natural language about some given image. Its complexity is based on the type of questions asked about the image. The type of questions can either be purely based on the content of the image or may require some extra information apart from the image. While lot of work has been done with purely image-content based questions in the past, some progress has also been made in the questions which require *common-sense*. In this project, we will explore VQA considering the questions which require *world-knowledge* that is represented with help of Knowledge Graphs (KG).

1 Introduction

Visual Question Answering (VQA) is an emerging problem ((Antol et al., 2015); (Malinowski et al., 2015)) concerning Computer Vision, Natural Language Processing and Artificial Intelligence where given an image and a natural language question (e.g.,*"What time of day it is?"*, *"Is that a frisbee?"*), the task is to automatically produce an accurate natural language answer ("night", "yes"). For this, it requires elements of natural language processing, image analysis, and text - visual entity linking. Set of questions that VQA models can answer exhibit some characteristics (e.g., counting, ordering, spatial, color) and form it's one of the limitations.

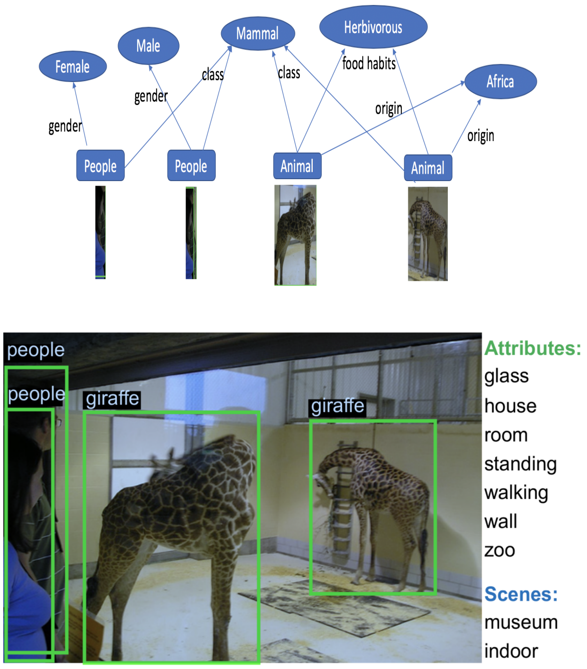
A number of works has been done recently in Visual Question Answering (VQA), but most of them exploit very basic interaction with the image only and do not require any external knowledge. They ask very simple questions, like *"What object is this?"* or *"It is of what colour?"* which are directly based on the content of the image. Some progress is also seen in working with *common-sense* questions, like *"Does it appear to be rainy?"* or *"Is this person expecting company?"*. But when we consider asking *knowledge-aware* questions which requires external information apart from the image, like *"who is the person?"* or *"What is his profession?"*, not much progress has been made.

Unlike regular VQA problem, *Knowledge-based Visual Question Answering* deals with answering question which requires *common-sense* knowledge involving *common nouns* (e.g., table, horse, etc) and *world knowledge* about *named entities* present in the image. This problem can be easily expanded to other tasks and play a significant role in human-machine interaction and medical assistance. However, it becomes difficult when AI systems are trained to learn and extract both language and vision content and encode *common-sense* knowledge and then to make reasoning by combining information from multiple sources to obtain the accurate answer. Inspite of gaining remarkable improvements in the performance of VQA models by using attention networks and multimodal embedding methods (Lu et al., 2018), the KVQA problem lies far away from achieving state-of-the-art results.

Dealing with *knowledge-aware* questions with respect to images is a new task and we are interested in exploring this field. Recently a dataset KVQA (Sanket Shah and Talukdar, 2019) is also introduced which comprises of Knowledge Graphs having background information about entities in images taken from Wikipedia.

2 Problem Statement

This motivates us to conduct comprehensive experiments on KVQA dataset, demonstrating and anlyzing the performance of existing VQA models on various parameters concerning attention mechanisms, visual question-related relation facts, face identification methods and question answering module. We will also try to provide methods to enhance the performance on this dataset.



Visual Question: How many giraffes are there in the image?
Answer: Two.

Common-Sense Question: Is this image related to zoology?
Answer: Yes. Reason: Object/Giraffe --> Herbivorous animals --> Animal --> Zoology; Attribute/Zoo --> Zoology.

KB-Knowledge Question: What are the common properties between the animal in this image and zebra?
Answer: Herbivorous animals; Animals; Megafauna of Africa.

Figure 1: Example taken from KB-VQA dataset (Wang et al., 2015) and the results given by Ahab (Wang et al., 2015), the proposed VQA approach. This example shows the variation between normal question, *common-sense* questions and *knowledge aware* questions.

3 Related Work

The initial work done in the field of VQA using external knowledge was done by Wang et al. 2017 (Wang et al., 2015). They came up with a method Ahab, which maps an image and its question to a query that is further applied to large scale structured Knowledge Base (KB) to get the final answer. The number of ways in which queries can be applied on knowledge base in this method are limited. They also proposed a dataset KB-VQA. The questions in this dataset, require external knowledge to answer. The evaluation protocol they used requires human evaluation, which is one more limitation apart from constraint on interaction with knowledge base.

Fact-based Visual Question Answering (Wang et al., 2016) provides deeper image understanding by modeling the set of questions that are answered using external source of information, such as Wikipedia. In contrast to the regular VQA datasets, FVQA dataset includes a supporting-fact for each question-answer pair for an image. The dataset is examined by measuring the performance of the state-of-the-art RNN (Recurrent Neural Network) based approaches. They propose a method to learn the mapping between questions and a set of KB-queries so that the supporting fact for reasoning can be provided, which is then used to form the final answer of the question. Their approach achieves the Top-1 accuracy of 56.91%, outperforming exixting baseline VQA methods (SVM (Cortes and Vapnik, 1995), LSTM (Hochreiter and Schmidhuber, 1997), Hie-Question-Image (Lu et al., 2016), Ensemble, and Human).

Recently a novel dataset - KVQA has been introduced by Sanket et. al. 2019 (Sanket Shah and Talukdar, 2019), which we are using for the sake of our project and is explained in detail in next section. Apart from the dataset, they have used memNet (Weston et al., 2014) as one of their baseline models. It has divided the task into following modules.

1. Entity Linking: Given an input image, caption and a question, entities present in the image and question are identified.

2. Fetching facts from KG: Knowledge facts are extracted in this module by traversing KG using the entities obtained in the previous module.

3. Memory and question representation: Each knowledge fact is fed to BLSTM to get corresponding memory embeddings. In this module, representation for output is calculated using attention over knowledge facts, question and memory embeddings.

4. Question answering module: It is the final module which produces the answer by applying softmax over output and question from previous module after feeding them into a multilayer perceptron.

4 Dataset and Metrics

4.1 Dataset

For our project we will be using KVQA dataset (Sanket Shah and Talukdar, 2019). It contains 183,007 question-answer pairs focusing on 18K persons present in 24,602 images. Questions in this dataset require multi-entity, multi-relation and multi-hop reasoning over KG to arrive at an answer. It also consists of ground-truth answers for the questions which go beyond KG entities.

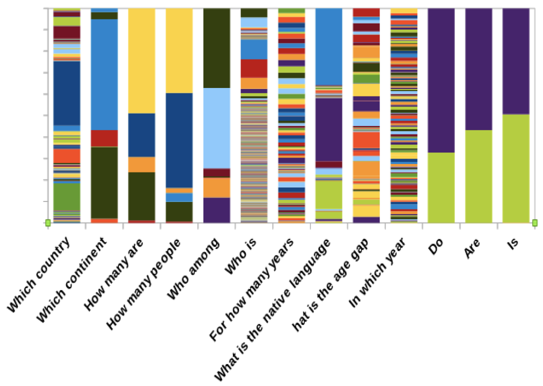


Figure 2: Analysis of question-answer pairs in the KVQA dataset. Distribution of answers for a few selected question types. Different colors are used to indicate different answers. For example, for questions starting with Do, Are and Is, the probable answers Yes and No are shown in green and purple, respectively (see the rightmost three bars).

KVQA dataset statistics:

Number of images	24,602
Number of QA pairs	183,007
Number of unique entities	18,880
Number of unique answers	19,571
Average question length (words)	10.14
Average answer length (words)	1.64
Average number of questions per image	7.44

Table 1: KVQA dataset statistics in brief.

They have also randomly divided 70%, 20% and 10% of images into train, test and validation respectively for comparison of VQA methods. This dataset contains 17K, 5K and 2K images, and corresponding approximately 130K, 34K and 19K question-answer pairs in one split of train, validation and test, respectively.

For knowledge graph, they have used RDF dump (dated: 05- 05-2018) of Wikidata (Vrandecic et al. 2014) . In this KB, knowledge is represented by a large number of triples of the form (arg1,rel,arg2), where arg1 and arg2 denote two entities in the KB and rel is a predicate representing the relationship between them. A collection of such triples form a large interlinked graph.

They also describe following two settings they used to evaluate their baselines, and we will be using the same.

1. Closed-world setting: In this setting, they use 18K entities and knowledge facts up to 3 hops away from these entities. Facts corresponding to 18 pre-specified relations are only considered. A few examples of such relations are occupation, country of citizenship, place of birth, etc.

2. Open-world setting: This relates to setting in the practical world. Here, a person in an image can be one of the 69K entities. Further, knowledge facts (up to 3-hops) from these entities constructed from 200 most frequent relations are used for reasoning over KG.

4.2 Metrics

We will evalualate our model on Top-N accuracy to compare the results obtained with the performance of the existing baseline methods on the KVQA dataset. Top-N accuracy means that the correct answer gets in the Top-N probabilities for the corresponding question-image(QI) pair, for it to count as correct.

5 Baselines

We are considering two baseline models: Memnet (Weston et al., 2014) and BiLSTMs.

1. Memnet: A memory network consists of a memory *m* and four parts I, G, O and R.

   I (input feature map): Converts the incoming input to the internal feature representation.

   G (generalization): Updates old memories given the new input.

   O (output feature map): Produces a new output, given the new input and the current memory state.

   R (response): Converts the output into desired the response format. For example, a textual response or an action.

2. BiLSTMS: We select BLSTMs as the second baseline to compare results (Top-1 accuracy and Top-5 accuracy) with that of our model. Memory embeddings are obtained by using each knowledge and spatial fact as input to BLSTM. Question embedding for a query is also obtained in the similar manner. It should be noted that multi-hop facts are given as sequential inputs to BLSTMs by ensuring that 1-hop facts occur before 2-hop facts and so on. This strategy to remember the sequence of occurance of facts is similar to MemNet.

6 Experimental Hypothesis

In KVQA paper, the baseline models under consideration were also Memnets and BLSTMS. Few observations in this work are:

- BLSTMs are ineffective when number of facts increase and memNet outperforms BLSTMs on all settings.
- Performance goes down by more than 10% when going from closed world to open world.
- memNet is inadequate in handling spatial, multi-hop, multi-relational, subtraction, counting and multientity questions, while it performs well on 1-hop, Boolean and intersection questions.

As we can see even memNet is performing well in only three categories and poor in rest of the categories, there is a clear need of question-guided approach for this task where better techniques can be designed based on question type.

Further they have mentioned two challenges they were facing. Firstly, face-descriptors like *Facenet* were not able to handle large number of distractors and secondly, question-answering technique was not able to scale up well with large number of *world-knowledge* facts associated with multiple entities.

We would like to focus on improving question-answering module for KVQA dataset. We do not propose any architecture now, but with the course of time we would like to conduct many experiments with standard VQA models considering KVQA dataset, and will try to further improve the performance currently achieved by baselines.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. *CoRR*, abs/1505.00468.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Sepp Hochreiter and Jrgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. *CoRR*, abs/1606.00061.

Pan Lu, Lei Ji, Wei Zhang, Nan Duan, Ming Zhou, and Jianyong Wang. 2018. R-VQA: learning visual relation facts with semantic attention for visual question answering. *CoRR*, abs/1805.09701.

Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. 2015. Ask your neurons: A neural-based approach to answering questions about images. *CoRR*, abs/1505.01121.

Naganand Yadati Sanket Shah, Anand Mishra and Partha Pratim Talukdar. 2019. Kvqa: Knowledge-aware visual question answering. In *AAAI*.

Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony R. Dick. 2015. Explicit knowledge-based reasoning for visual question answering. *CoRR*, abs/1511.02570.

Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony R. Dick. 2016. FVQA: fact-based visual question answering. *CoRR*, abs/1606.05433.

Jason Weston, Sumit Chopra, and Antoine Bordes. 2014. Memory networks. *CoRR*, abs/1410.3916.