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 and Problem Statement

1.1 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.

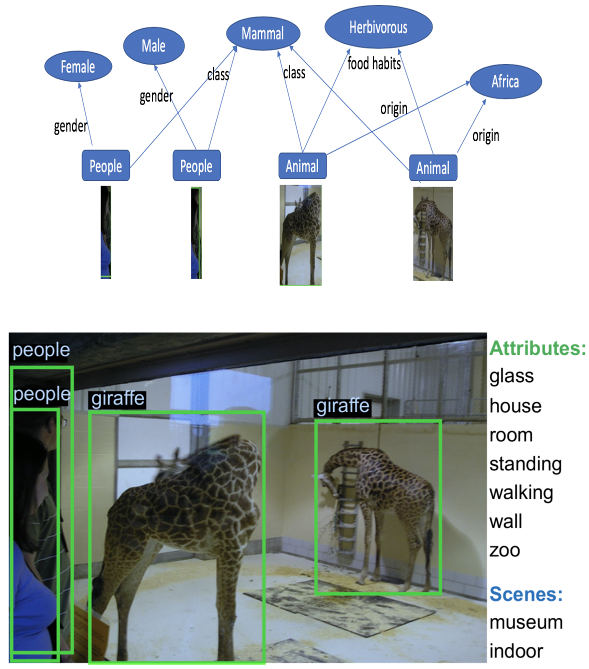
A number of works has been done recently in Visual Question Answering (VQA), but most of them exploit very basic interaction with the image only and do not require any external knowledge. They ask very simple questions, like *"What object is this?"* or *"It is of what colour?"* which are directly based 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.

1.2 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.

2 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.

3 Knowledge-aware VQA Network (KVQAN)

In the following, we first provide an overview of the proposed architecture KVQAN for knowledge based visual question answering before discussing our embedding space and learning formulation.

3.1 Interaction with KVQA dataset

We are considering the dataset D provided in which named entity part is already solved and named entities are provided along with the question instead of the image. So, we have firstly, data D in which questions Q_1, Q_2,.., named entities ne_1, ne_2,.. and answers A_1, A_2,.. (naming only the parts we considered in the project) are present and secondly the knowledge graph KG which contains all facts.

In the KG they have used code in the place of names of people, and we have also solved that part at various places, but that is not much relevant and we are skipping explaination of that part for simplicity.

3.1.1 About Modules

There are mainly 5 modules but all the processing is not done in the modules itself. The flow through the architecture will be explained in the next sub section.

1. **Fact Module:** It provides embedding of the facts passed to it. It simply replaces the words by their GloVe embedding, and then average them to get final embedding of that fact. This is a one time thing, and no training is involved in this part. Original embeddings are denoted by F_1, F_2,.. and embeded facts are by f_1, f_2,...

2. **Question Module:** This module provides embedding for the question Q, which is q , which we get through a GRU. We use GloVe embedding for initializing the embeddings to get it more closer to the facts embeddings. Although we have set the trainable parameter for embeddings to be true.

3. **Memory Module:**

   The memory module, retrieves information from 4 most relevant facts f_1, f_2, f_3, f_4 for a question q provided to it by focusing attention on these facts. We will be having 4 most relevant facts f_1, f_2, f_3, f_4 for a question q, which we have got scoring all facts using cosine similarity. These facts will be input of this module and from output we desire an embedding by these four facts, which we call f_{final}. We implement this attention by associating a single scalar value, attention g_i^t with each fact $\bar{f}_i$ during pass t. This is computed by allowing interactions between the fact and both the question representation and the memory state.

$$z_i^t = [\bar{f}_i \circ q; \bar{f}_i \circ m^{t-1}; |\bar{f}_i - q|; |\bar{f}_i - m^{t-1}|]$$

$$Z_i^t = W^{(2)} tanh(W^{(1)} z_i^t + b^{(1)}) + b^{(2)}$$

$$g_i^t = \frac{exp(Z_i^t)}{\sum_{k=1}^{M_i} exp(Z_k^t)}$$

   where $\bar{f}_i$ is the i^{th} fact, m^{t-1} is the previous memory, q is the original question, $\circ$ is the element-wise product, $|.|$ is the element-wise absolute value, and ; represents concatenation of the vectors.

4. **Linear Layer Module:** We concatenate question embedding q and final fact embedding f_{final} so it will give a 200 dimension vector. It will be passed into a linear layer and its result will be a 100 dimensional embedding or vector (q+f).

5. **Answer Module:** Now the (q+f) embedding is passed through a GRU to get the final answer a. We have used a GRU because answers are multi-words. If they would have been single word, a simple linear layer would have been sufficient.

4 Dataset and Metrics

4.1 Dataset

We use the publicly available KVQA dataset (Sanket Shah and Talukdar, 2019). and its knowledge base to evaluate our model. It consists of 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 contains of ground-truth answers for the questions which go beyond KG entities.

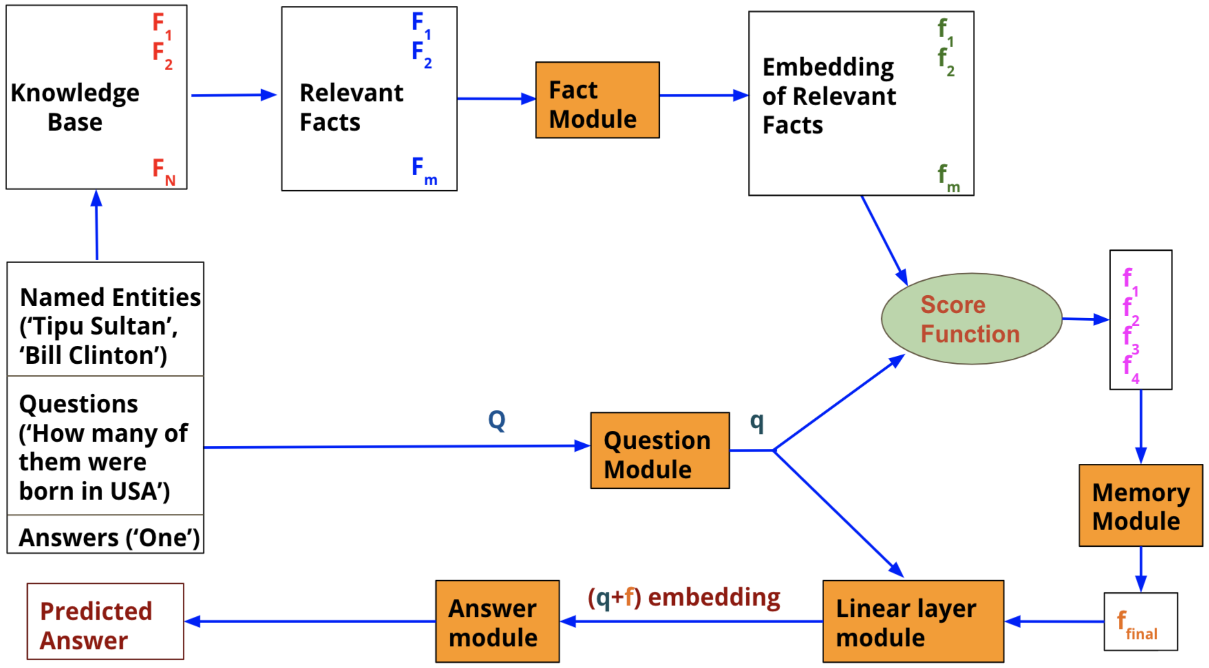


Figure 2: KVQAN Architecture

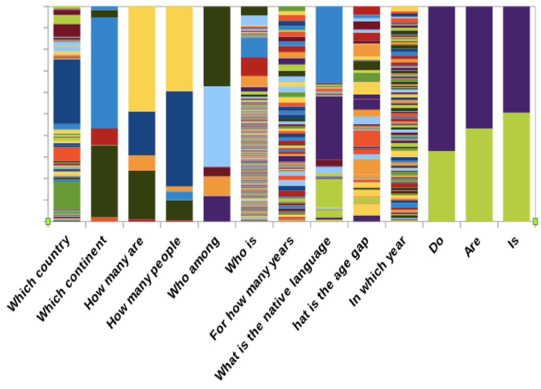


Figure 3: Analysis of question-answer pairs in the KVQA dataset. Distribution of answers for a few selected question types. Different colors are used to indicate different answers. For example, for questions starting with Do, Are and Is, the probable answers Yes and No are shown in green and purple, 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.

The dataset also provides random splits of images into train (70%), test (20%) and validation (10%) sets for comparison of VQA methods. It 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, RDF dump (dated: 05-05-2018) of Wikidata (Vrandecic et al. 2014) was used. 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 original work (cite KVQA) describe two settings to evaluate their baselines, and we have used 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 im-

age 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 evaluate our model on accuracy to compare the results obtained with the performance of the existing baseline methods on the KVQA dataset. For ans answer to count as correct, we use exact string matching.

In the KVQA paper, they have not well specified anything about metric they haves used, and just shown numbers. We have assumed that to be accuracy and therefore used it.

5 Baselines

We have considered 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 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 occurrence of facts is similar to MemNet.

6 Experiments

1. **Getting relevant facts:** We searched for named entities in the entire knowledge graph, and extract all the facts which contains the named entities.

2. **Getting embedding of relevant facts:** After getting the relevant facts we passed them through fact module to get their embedding. It is done by replacing each word with its GloVe embedding and then taking average of all.

3. **Getting Question embedding:** Question embedding is got by passing question through a GRU layer, but initialized by GloVe embedding before passing it into GRU layer. The trainable parameter is set to True.

4. Now, we have both embeddings for question and relevant facts. Now we will extract four most relevant facts by scoring them against question embedding. It is done by simply taking cosine similarity between fact's embedding and question embedding, and later we got four embedding which had the top most score.

5. Now, embedding of fact and question are available, we have concatenated them and passed through a linear layer to get a mixed representation of question and fact.

6. **Predicting Answers** Now the representation of fact and question is passed through a GRU and answer is predicted

7. **Experimental settings** We have considered negative log-likelihood loss. Optimizer used is Adam with learning rate 0.001. We trained the model for 14 epochs, and 1 epoch took around 50 minutes. Size of embedding is 100, and every where hidden size is also 100. Maximum fact length is 120, question length is 92 and answer length is 12. Padding is done accordingly. For training we have considered initial 20,000 data points and rest which are around 4000 are used for testing.

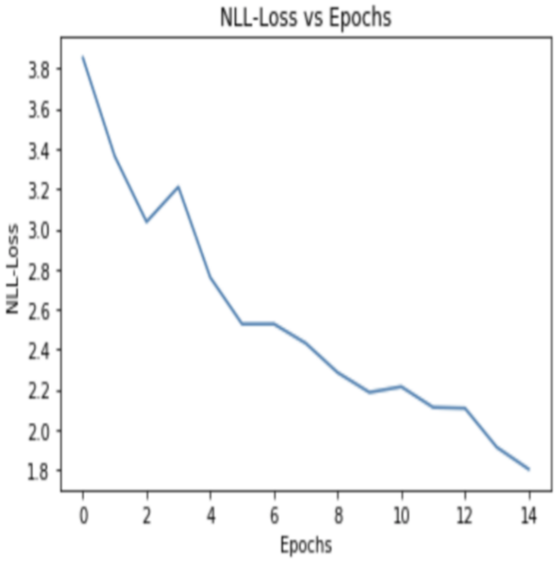


Table 2: Loss vs Epochs

7 Results

In the following, we assess the proposed approach. We first provide details about the quantitative results for prediction of relevant facts from knowledge base, prediction of answer-source from questions, and prediction of the answer and the supporting fact. Next, we provide the qualitative results.

We are working in open-world setting and additionally in the KVQA paper, they have not provided any information about closed world setting they have used.

Apart from that we are assuming oracle setting, in which they were saying that facial recognition is assumed to be solved. But they had both named entity and its coordinates in the image but we only have named entities and no information about its spatial position in the image.

The accuracy we got on test set is 29.223. In the paper, they have not explored this setting (Oracle + open-world). But if we compare our score with their open-world, no wikipedia captions and original questions setting, their BLSTM score is 16.8 and MemNet score is 36.0

8 Discussion

Our results suggest that despite progress in question answering, significant further research is needed to achieve high performance knowledge-based visual question answering.The highly successful face recognition systems such as FaceNet fails to address challenges due to large number of distractors. Similarly, question answering technique such as memory networks (Weston, Chopra, and Bordes 2014) does not scale well with large number of world knowledge facts associated with multiple entities.

9 Future Work

In this work, we addressed knowledge-based visual question answering and propose a method that learns to embed facts as well as question-image pairs into a space that admits efficient search for answers to a given question. In contrast to existing retrieval based techniques, our approach learns to embed questions and facts for retrieval. In the future, we hope to address extensions of our work to larger structured knowledge bases, as well as unstructured knowledge sources, such as online text corpora.

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.