KVQA: Knowledge-Aware Visual Question Answering

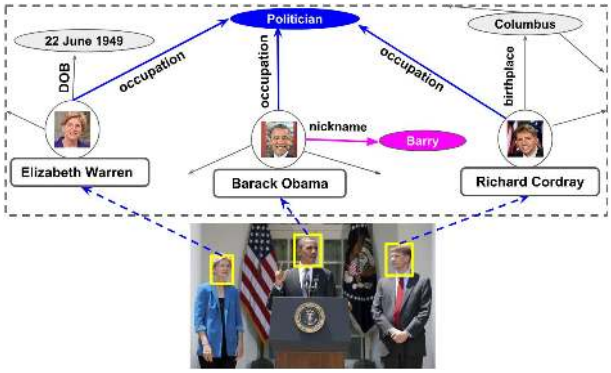
Sanket Shah,[1*] Anand Mishra,[2*] Naganand Yadati,[2] Partha Pratim Talukdar[2]
[1]IIIT Hyderabad, India, [2]Indian Institute of Science, Bangalore, India
sanket.shah@research.iiit.ac.in, anandmishra@iisc.ac.in, naganand@iisc.ac.in,ppt@iisc.ac.in
[1,2]: contributed equally to the paper.
[1]: Research carried out during an internship at the Indian Institute of Science, Bangalore.

Abstract

Visual Question Answering (VQA) has emerged as an important problem spanning Computer Vision, Natural Language Processing and Artificial Intelligence (AI). In conventional VQA, one may ask questions about an image which can be answered purely based on its content. For example, given an image with people in it, a typical VQA question may inquire about the number of people in the image. More recently, there is growing interest in answering questions which require *commonsense knowledge* involving *common nouns* (e.g., cats, dogs, microphones) present in the image. In spite of this progress, the important problem of answering questions requiring *world knowledge* about *named entities* (e.g., Barack Obama, White House, United Nations) in the image has not been addressed in prior research. We address this gap in this paper, and introduce KVQA – the first dataset for the task of (world) knowledge-aware VQA. KVQA consists of 183K question-answer pairs involving more than 18K named entities and 24K images. Questions in this dataset require multi-entity, multi-relation, and multi-hop reasoning over large Knowledge Graphs (KG) to arrive at an answer. To the best of our knowledge, KVQA is the largest dataset for exploring VQA over KG. Further, we also provide baseline performances using state-of-the-art methods on KVQA.

Introduction

We are witnessing a strong wave in Artificial Intelligence. Artificially intelligent personal assistants, chatbots, and robots are becoming reality for specific tasks. In this context, Visual Question Answering (VQA), which given an image aims to answer questions, provides an excellent tool to communicate effectively with such artificial agents. There has been noticeable progress in VQA in the past few years (Antol et al. 2015; Goyal et al. 2017; Trott, Xiong, and Socher 2018; Su et al. 2018). Questions in conventional VQA can usually be answered purely based on the content of the image alone. More recently, there is also growing interest to answer visual questions which require *commonsense knowledge*, i.e., knowledge about *common nouns*[1] (Wang et al. 2017; 2018; Su et al. 2018; G. Narasimhan and Schwing 2018). An example of such commonsense knowledge is *microphone is used for amplifying sound*, where *microphone* is a common noun. Examples of a few conventional as well as



Conventional VQA (Antol et al. 2015; Goyal et al. 2017; Trott, Xiong, and Socher 2018)
Q: How many people are there in the image?
A: 3

Commonsense knowledge-enabled VQA (Wang et al. 2017; 2018; Su et al. 2018; G. Narasimhan and Schwing 2018)
Q: What in the image is used for amplifying sound?
A: Microphone

(World) knowledge-aware VQA (KVQA, this paper):
Q: Who is to the left of Barack Obama?
A: Richard Cordray
Q: Do all the people in the image have a common occupation?
A: Yes
Q: Who among the people in the image is called by the nickname Barry?
A: Person in the center

Figure 1: An example image, questions and answers (shown in blue) in three settings: (i) Conventional VQA, (ii) Commonsense knowledge-enabled VQA, and (iii) (World) knowledge-aware VQA, which requires knowledge and reasoning involving the *named entities* present in the image (e.g., Barack Obama, Richard Cordray, etc. in the above image). *[Best viewed in color].*

[1]common noun: a noun representing a class of objects or a concept, as opposed to a particular individual. (Oxford Dictionary)

commonsense-enabled VQA are shown in Figure 1.

However, in many real-world situations of interest, we also need to satisfy our information need about the *named entities*[2] present in images. For example, given the image in Figure 1, we would like to answer questions such as *Who is to the left of Barack Obama?*, or *Do all the people in the image have a common occupation?* Answering such questions require *world knowledge* about the named entities present in the image, and also reason over such knowledge. We refer to this problem as knowledge-aware VQA (KVQA). Despite having many real-world applications, this problem has not been explored in literature, and existing datasets as well as methods are inadequate. This calls for the need of a new dataset to initiate research in this direction.

Wikidata (Vrandecic and Krötzsch 2014) is a large-scale collaboratively edited Knowledge Graph (KG) that contains 50 million items including *world knowledge* about persons, taxons, administrative territorial entities, architectural structure, chemical compound, movies, etc. In context of such freely-available web-scale knowledge bases, answering questions which require background knowledge about the named entities appearing in an image is a challenging task. The key challenges associated with this task are: (i) identifying all the named entities at scale. It should be noted that number of common nouns (objects) is way smaller than number of ever-growing proper nouns (named entities) in the real world, (ii) mining relevant knowledge facts from web-scale KG, and (iii) learning to reason from the mined facts. While addressing all these challenges is beyond the scope of a single paper, we make the first step towards this, and introduce a dataset KVQA containing images of persons and manually verified question-answer pairs related to the background knowledge about them.

Our dataset KVQA can be viewed and downloaded from our project website: http://malllabiisc.github.io/resources/kvqa/. It contains 183K question-answer pairs about more than 18K persons and 24K images. The questions in this dataset require multi-entity, multi-relation and multi-hop reasoning over KG to arrive at an answer. To the best of our knowledge, KVQA is the largest dataset for exploring VQA over KG. The Visual Question Answering on our dataset naturally leads to the problem of visual named entity linking where the task is to link the named entity appearing in an image to one of the entities in Wikidata. To enable visual named entity linking, we also provide a support set containing reference images of 69K persons harvested from Wikidata as part of our dataset. The visual named entity linking task in this paper could lead to one of the largest one-shot classification problems in Computer Vision. Further, we provide baselines using state-of-the-art methods for visual named entity linking as well as Visual Question Answering on KVQA.

The contributions of this paper are as follows.

1. We draw attention to the important but unexplored problem of Visual Question Answering (VQA) involving *named entities* in an image. Answering named entity-focused questions require knowledge about the world and reasoning over it, and hence we term this new task as *knowledge-aware VQA* (KVQA).
2. We introduce KVQA, the first dataset for the proposed (world) knowledge-aware VQA task. KVQA is more than 12x larger than recently proposed commonsense-enabled VQA datasets. Please note that KVQA is the only dataset which recognizes named entities and the need for knowledge about them.
3. We also provide performances of state-of-the-art methods when applied over the KVQA dataset. We are optimistic that KVQA will open up various new research avenues spanning Computer Vision, Natural Language Processing, and more broadly AI.

[2]named entity: a thing with distinct and independent existence (Oxford Dictionary)

Related Work

Question answering about image, also popularly known as Visual Question Answering (VQA), has gained huge interest in recent years (Goyal et al. 2017; Antol et al. 2015; Zhu et al. 2016; Kembhavi et al. 2017). The recent interest on this problem is primarily due to significant improvement in image understanding. Moreover, introduction of large-scale Visual Question Answering datasets, such VQA (Antol et al. 2015) and VQA v2.0 (Goyal et al. 2017) has played crucial role in VQA research. One of the earliest VQA databases viz. DAQUAR (Malinowski and Fritz 2014) contains only 12K question-answer pairs about images with a potential answers as object names, colors, or their combination. Similarly, (Geman et al. 2015) consider template based questions generated from a fixed vocabulary containing object names, object attributes and relationships between objects. The introduction of VQA benchmarks[3] has significantly changed the research in this area. Recently, there has also been interest in bringing compositional language and elementary visual reasoning for Visual Question Answering (Johnson et al. 2017). However, they are still limited to reasoning over synthesized objects in image alone, and not over named entities present in image and external knowledge both as necessary in KVQA.

Recognizing the need of external knowledge in answering questions about a given image, commonsense-enabled VQA is a recent interest of the community (Wang et al. 2018; 2017; Su et al. 2018; G. Narasimhan and Schwing 2018; Aditya, Yang, and Baral 2018). In this context, two datasets, namely KB-VQA (Wang et al. 2017) and FVQA (Wang et al. 2018) were recently introduced. However, these datasets are centered around common nouns, and contain only a few images, just 700 images in KB-VQA and 1906 images in FVQA. Further, they are far more restricted in terms of logical reasoning required as compared to the dataset we introduce in this paper. Other recent datasets such as (Kembhavi et al. 2017; Tapaswi et al. 2016; Lu et al. 2018) do not require external knowledge as ours. The summary of datasets and comparison with ours is presented in Table 1.

Computer Vision and Knowledge Graph: There is a growing interest in combining Computer Vision and back-

[3]http://www.visualqa.org/

Dataset name	# images	# QA pairs	Image source	Knowledge Graph type	Named entities
KVQA (this paper)	24,602	183,007	Wikipedia	World knowledge	✓
FVQA (Wang et al. 2018)	1,906	4,608	COCO	Commonsense	✗
KB-VQA (Wang et al. 2017)	700	2,402	COCO + ImgNet	Commonsense	✗
TallyQA (Acharya, Kafle, and Kanan 2019)	165,443	287,907	Visual genome + COCO	✗	✗
CLEVR (Johnson et al. 2017)	100,000	999,968	Synthetic images	✗	✗
VQA-2 (Goyal et al. 2017)	204,721	1,105,904	COCO	✗	✗
Visual Genome (Krishna et al. 2017)	108,000	1,445,322	COCO	✗	✗
TQA (Kembhavi et al. 2017)	-	26,260	Textbook	✗	✗
Visual 7w (Zhu et al. 2016)	47,300	327,939	COCO	✗	✗
Movie-QA (Tapaswi et al. 2016)	-	14,944	Movie videos	✗	✗
VQA (Antol et al. 2015)	204,721	614,163	COCO	✗	✗
VQA-abstract (Antol et al. 2015)	50,000	150,000	Clipart	✗	✗
COCO-QA (Ren, Kiros, and Zemel 2015)	69,172	117,684	COCO	✗	✗
DAQUAR (Malinowski and Fritz 2014)	1,449	12,468	NYU-Depth	✗	✗

Table 1: Comparison of KVQA (introduced in this paper) with various other previously proposed representative VQA datasets. This list is not meant to be exhaustive, nor to describe the datasets in detail, but merely to provide a sample of VQA datasets that are available. We note that KVQA is the only dataset which requires *world knowledge* about named entities present in the images to answer questions about them.

ground knowledge, usually Knowledge Graph. This interest is not just limited to applications such as VQA, but is also used for traditional Computer Vision tasks such as image classification (Marino, Salakhutdinov, and Gupta 2017), object detection (Fang et al. 2017), zero-shot image tagging (Lee et al. 2018). However, most of these works still focus on bridging images and commonsense Knowledge Graphs like WordNet (Miller 1995) and Conceptnet (Liu and Singh 2004), unlike ours which requires bridging images and world knowledge.

Other datasets: Recognizing and understanding person activities has been one of the main-stream research areas in Computer Vision. The problems in this research area includes, recognizing celebrities (Guo et al. 2016), human-pose estimation (Toshev and Szegedy 2014), action recognition (Simonyan and Zisserman 2014), recognizing celebrities in places (Zhong, Arandjelović, and Zisserman 2016), and large-scale attribute prediction in celebrity faces (Liu et al. 2015). In these works, images were treated separately from *world knowledge*, except MS-Celebs (Guo et al. 2016) which provides Freebase IDs associated with cropped faces of celebrities. However, even MS-Celebs can not be utilized for the task of asking questions about multiple entities and their spatial order in the image.

Face identification and context: The visual entity linking task, i.e., linking named entities appearing in the image to Knowledge Graph leads to a need for face identification at web scale. Face identification has been well-studied and highly successful problem in Computer Vision, and large-scale train sets for face identification have been introduced (Guo et al. 2016; Cao et al. 2018). However, in literature, testing is still performed on a significantly smaller scale than what KVQA requires. There have also been works demonstrating utility of context in improving face identification often in restricted settings (Bharadwaj, Vatsa, and Singh 2014; Lin et al. 2010; O'Hare and Smeaton 2009). We believe that our dataset can be used to further move forward such studies in a much more wilder setting.

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 2: KVQA dataset statistics in brief. We also provide train-validation-test splits and evaluation protocols.

KVQA

We introduce a novel dataset – KVQA – to study Visual Question Answering over Knowledge Graph. The dataset contains 183K question-answer pairs about 18K persons contained within 24K images. Questions in this dataset require multi-entity, multi-relation and multi-hop reasoning over KG to arrive at an answer. Further, questions which go beyond KG entities as ground-truth answers is another unique characteristics of our dataset. The brief statistics of our dataset is given in Table 2.

Data collection and annotation.

KVQA is annotated using the following stages.

S1– Image collection: We compile an exhaustive list of persons from Wikidata (Vrandecic and Krötzsch 2014) contain-

(a) ***Wikipedia caption*:** Khan with United States Secretary of State Hillary Clinton in 2009.

Q: Who is to the left of Hillary Clinton? (*spatial*)
A: **Aamir Khan**

Q: Do all the people in the image have a common occupation? (*multi-entity, intersection, 1-hop, Boolean*)
A: **No**

(b) ***Wikipedia caption*:** Cheryl alongside Simon Cowell on The X Factor, London, June 2010.

Q: What is the age gap between the two people in the image? (*multi-entity, subtraction, 1-hop*)
A: **24 years**

Q: How many people in this image were born in United Kingdom? (*1-hop, multi-entity, counting*)
A: **2**

(c) ***Wikipedia caption*:** BRICS leaders at the G-20 summit in Brisbane, Australia, 15 November 2014

Q: Were all the people in the image born in the same country? (*Boolean, multi-entity, intersection*)
A: **No**

Q: Who is the founder of the political party to which person second from left belongs to? (*spatial, multi-hop*)
A: **Syama Prasad Mookerjee**

(d) ***Wikipedia caption*:** Serena Williams and Venus Williams, Australian Open 2009.

Q: Who among the people in the image is the eldest? (multi-entity, comparison)
A: **Person in the left**

Q: Who among the people in the image were born after the end of World War II? (multi-entity, multi-relation, comparison)
A: **Both**

Figure 2: A selection of images and question-ground truth answer pairs from the KVQA dataset. Image also comes with captions from Wikipedia. Challenges associated with questions are shown in parenthesis.

ing athletes, politicians and actors. Once this list is prepared, we harvest images and their captions from Wikipedia pages of these persons. In this stage, we consider 30K persons and download around 70K images. Note that not all images in this set are useful for our task as we discuss next.

S2– Counting, identifying and ordering persons in image: Our stage-2 annotation is first to count the number of people in images. To this end, we provide images to a team of human annotators for grouping images based on the number of persons. At the end of this phase, we prune all images which either do not contain any person or are overly crowded. The reason for this pruning is because such images are not very useful towards asking meaningful questions against. We also remove all the exact duplicate images to avoid redundancy. We then ask human annotators to identify the person and give their correct ordering from left to right in the image. More specifically, we ask them to identify the persons in the image, their spatial order, and their corresponding Wikipedia pages to avoid disambiguation. As an aid, we provided Wikipedia captions along with images to the annotators.

S3– Obtaining ground-truth answers: Once we knew the named entities and their relative order in the image, we asked the annotators to give us a list of templated-questions which can be asked to know more about the persons in the image. Given these templated-questions and order of people in the image, we wrote SPARQL queries to obtain answers from Wikidata (Vrandecic and Krötzsch 2014). It should be noted that the major objective of our dataset is to promote development of end-to-end trainable VQA system, and hence the templates and SPARQL queries are only used to obtain ground truth, and are not provided along with the questions. We strongly recommend researchers to only use question-answer-image tuples along with KG to train their VQA system. Further, to make questions more challenging and realistic, we paraphrase them as discussed in step S5 (below).

S4– Training, validation, and testing: In order to facilitate fair comparison of VQA methods in future, we provide training, validation and testing splits. We randomly divide 70%, 20% and 10% of images respectively for train, test, and validation, respectively. KVQA 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. We provide five such splits for our dataset, and suggest that each experimenter, at a minimum, report mean accuracy computed over all the test splits.

S5– Paraphrasing questions: In order to increase the complexity of the task and to make it closer to the real-world, we also paraphrased the questions in the test set. We used an online tool[4] to get suggestions for paraphrased questions. These suggestions were provided to human annotators to either accept, refine or discard them. A couple of examples of question paraphrasing are given here: (i) Original question *"Who among the people in the image were ever married to Paula Ben-Gurion?"* is paraphrased to *"Who among*

[4]https://paraphrasing-tool.com/

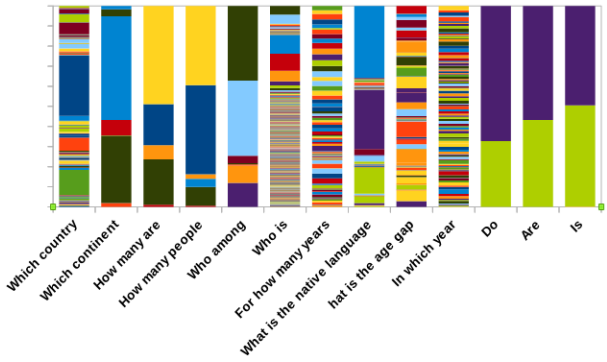


Figure 3: **Analysis of question-answer pairs in the KVQA dataset.** We show distribution of answers for a few selected question types. Different colors are used to indicate different answers. For example, for questions starting with *Do*, *Are* and *Is*, the probable answers *Yes* and *No* are shown in green and purple, respectively (see the rightmost three bars). We observe that answers are more evenly distributed which implies the balance in our dataset. *[Best viewed in color].*

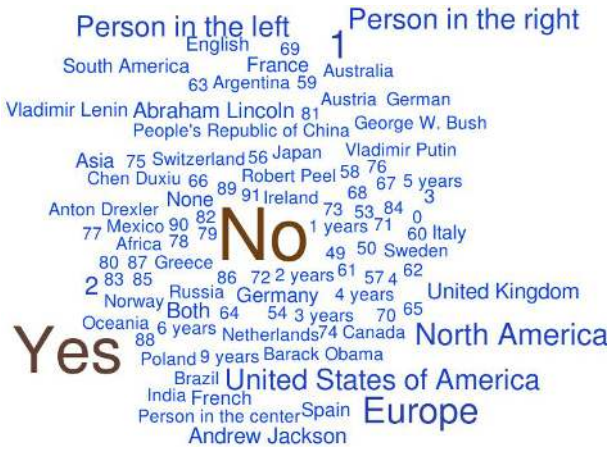


Figure 4: **Word cloud of answer words and phrases in the KVQA dataset.** We show word cloud for the top-100 most frequent words. We observe that there are a lot of variations in the answers, e.g., answer *Yes* and *No* are almost equally likely for Boolean questions in our dataset, whereas other answers are nearly uniformly distributed.

the people in the picture at any point wedded to Paula Ben-Gurion?". (ii) Original question *"Were all the people in the image born in the same continent?"* is paraphrased to *"Were all the folks in the picture took birth in the same landmass?"*. We evaluate baseline VQA approaches on both original and paraphrased questions.

Analysis of KVQA

We observe that KVQA contains challenges for both the vision community, e.g., handling faces with different poses, race and scale, and large coverage of persons, as well as for the language understanding community, e.g., answering questions which require multi-hop and quantitative reasoning over multiple entities in Knowledge Graph. We also group questions in our dataset based on challenges into the following ten categories.

Question challenge categories: *spatial, 1-hope, multi-hop, Boolean, intersection, subtraction, comparison, counting, multi-entity, and multi-relation*

Note that these categories are not mutually exclusive in our dataset, and a question can pose more than one challenge. We show a few examples of images and question-answer pairs in Figure 2 with associated challenges in parenthesis. For example, consider Figure 2(c). One question for this image of BRICS leaders is *"Who is the founder of the political party to which person second from left belongs to?"* requires understanding of "what is second from left" along with who the people are in the image, and multi-hop *world knowledge* facts like *Narendra Modi is a member of BJP*, and *BJP is founded by Syama Prasad Mookerjee*.

We further analyze the distribution of answers in the KVQA dataset. Biased and unbalanced answers in VQA datasets has been a challenge, and often blindly answering without understanding image content leads to a superficially high performance. For example, as studied by (Goyal et al. 2017), in the popular VQA (Antol et al. 2015) dataset, questions starting with *"What sport is"* very often have *"Tennis"* as the answer. Similarly, in such datasets most often (87%) Boolean questions have "Yes" as answer. We avoid this by asking diverse questions intelligently. In order to demonstrate the diversity in answers for various questions in the KVQA dataset, we show the distribution of answers for a few question types in Figure 3. Moreover, to illustrate presence of many unique answers in our dataset, we illustrate word cloud of top-100 most frequent answer words and phrases in Figure 4. We observe a lot of variations in answers indicating the challenging nature of the KVQA dataset.

Knowledge Graph

We use Wikidata (Vrandecic and Krötzsch 2014) as our Knowledge Graph. Note that Wikidata is a collaboratively edited knowledge base hosted by the Wikimedia Foundation. Specifically, we used the latest RDF dump (dated: 05-05-2018) of this KG. It stores facts in the form of triples, i.e., subject, relation and object. Each entity and relation are represented as unique Qids and Pids, respectively. For example, *Barack Obama* is represented as *Q76* and relation *has-nickname* is represented as *P1449*. This KG originally contains 5.2K relations, 12.8M entities and 52.3M facts. We consider a subset of these fact dumps for 114K persons. This list of persons is union of entities present in MS-Celebs (Guo et al. 2016) and KVQA. We harvest reference image for these persons from Wikidata. However, due to incompleteness of KG, we could only obtain reference images for 69K persons. These reference images are used for visual entity linking in one-shot setting as we discuss in the next section. All these preprocessed data and reference images are publicly available in our project website.

We now describe two settings in which we evaluate our baselines.

Method	Precision		Recall	
	Top-1	Top-5	Top-1	Top-5
MS-Captionbot	95.2	NA	16.6	NA
Facenet [Closed world]	81.0	-	82.2	87.2
Facenet [Open world]	73.5	-	71.4	76.5

Table 3: Results of Visual Entity Linking which is equivalent to face identification over KVQA. Please refer to the "Visual Entity Linking" section for more details.

1. **Closed-world setting:** In this setting, we 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 is a much open and practical setting. 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.

Approach

Visual Entity Linking.

Analogues to entity linking (Shen, Wang, and Han 2015) in NLP where the task is to link entity mentions in text to KG, the visual entity linking problem aims to link visual entities to KG. In our KVQA dataset, visual entities are persons. Hence, for us, visual entity linking becomes equivalent to face identification – however at web scale, ideally covering millions of entities. While it is an ambitious goal of our project, we have done a running start towards it, and created a benchmark for visual entity linking where person faces in images have to be linked to one of the 69K entities in KVQA.

There have been recent attempt to address face identification at large scale. However, most of the previously proposed face identification benchmarks (Cao et al. 2018; Guo et al. 2016) contain multiple reference images per person. Further, they restrict themselves to cover smaller set of unique persons during test time. Contrary to this, we advocate testing at larger scale, and with support of just a few reference images (as few as one).

Face identification at scale: Face localization is a precursor to face identification in our dataset. We explore two successful modern methods (Zhang et al. 2016; Hu and Ramanan 2017), and choose a better performing method (Zhang et al. 2016) for face localization. Once face localization is done, our next step is to do face identification. We use two baselines for this task: MS-Captionbot[5] and Facenet (Schroff, Kalenichenko, and Philbin 2015). MS-Captionbot is designed to provide captions for images. However, it also recognizes a few selected public figures. We use the generated caption and extract the named entities from it using an open-source entity linker, Dexter[6]. Facenet is a deep architecture designed to obtain robust face representation. The pretrained model of Facenet is publicly available, and regularly updated with training performed on larger benchmarks. We use pretrained models released on 2017-05-12 for face identification. It should be noted that Facenet achieves near perfect face verification performance on recent benchmarks including VGGFace2 (Cao et al. 2018).

We obtain face representation for reference images as well as faces localized using (Zhang et al. 2016) for our dataset, and perform 1-nearest neighbor to identify faces as one of the 18K (in closed-world) or 69K (in open-world) named entities. Visual entity linking results on KVQA is reported in Table 3. We observe that MS-Captionbot achieve high precision for visual entity linking task, but its recall is poor. Note that higher recall is desired for a successful KVQA system. Based on this, we choose (Schroff, Kalenichenko, and Philbin 2015) for visual entity linking in subsequent task, i.e., VQA over KG. However, we also notice that off-the-shelf face descriptor (Schroff, Kalenichenko, and Philbin 2015) does not deliver well in open-world setting indicating challenges associated with face identification at scale.

VQA over KG

Once visual entities are linked to Wikidata, the task of Visual Question Answering becomes equivalent to fetching relevant facts from KG, reasoning over them, and learning to answer questions. One may also design end-to-end system where visual entity linking gets benefited from question-answer pairs. However, in this work, we assume visual entity linking and question answering as two separate modules.

We choose memory network (memNet) (Weston, Chopra, and Bordes 2014) as one of our baselines for KVQA. Memory network provides a general architecture for learning from external knowledge. Note that memNet and their variants show state-of-the-art performance in VQA task (Su et al. 2018) requiring commonsense knowledge. The modules of our memory network which is also illustrated in Figure 5, are as follows:

1. **Entity linking**: Given an input image, corresponding Wikipedia caption (which is optional for us) and a question, our first goal is to identify entities present in the image and question. As discussed earlier, visual entities in our datasets are persons in Wikidata. Hence, we identify them as discussed in the previous section with the help of reference images obtained from Wikidata. In order to obtain entities present in text, i.e., question and Wikipedia caption, we run an open-source named entity recognizer (viz. Dexter). The visual entity linking and text entity linking are separately shown in Figure 5. At the end of this stage, we obtain list of entities present in image and text. Additionally, for visual entities we also obtain their spatial position, i.e., face co-ordinates in image.
2. **Fetching facts from KG**: Once we obtain list of entities from above module, we traverse KG from these entities

[5]https://www.captionbot.ai/

[6]http://dexter.isti.cnr.it/

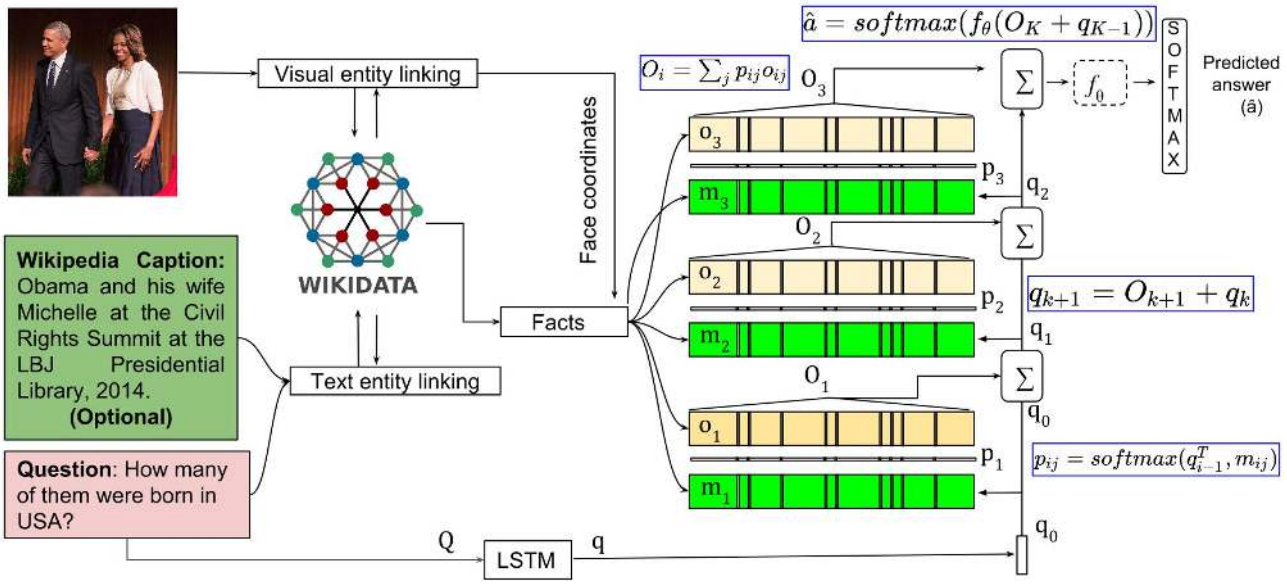


Figure 5: Memory Network-based state-of-the-art baseline for KVQA. Memory networks provides a general architecture for learning from external knowledge, and their variants show state-of-the-art performance in VQA task (Su et al. 2018) requiring commonsense knowledge. Please refer to the "VQA over KG" section for more details.

and extract knowledge facts. We restrict ourself to three-hops in this work. Further, we augment these knowledge facts with spatial facts such as *Barack Obama is at* (x_1, y_1) and *Michelle Obama is at* (x_2, y_2). Here (x_i, y_i) corresponds to center of face bounding boxes in image.

3. **Memory and question representation**: Each knowledge and spatial fact is fed to BLSTM to get corresponding memory embeddings m_i. Question embeddings (q) for a question (Q) is also obtained in a similar fashion.

   We then compute a match between q and each memory m_{ij} by taking the dot product followed by a softmax as follows.

   $$p_{ij} = \text{softmax}(q^T m_{ij}), \tag{1}$$

   where

   $$\text{softmax}(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)}.$$

   Here, p_{ij} acts as a soft attention over knowledge facts given question Q. Given these, we compute final output representation O by linearly combining output representations for all the facts as follows.

   $$O = \sum_j p_{ij} o_{ij}.$$

4. **Question answering module:** The sum of output representation O and question q are fed to a multi-layer perceptron f_θ where θ is a trainable parameter. Then, a softmax over output is used to predict answer $\hat{a}$.

   $$\hat{a} = \text{softmax}(f_\theta(O + q)).$$

During training, we train memory and question representation and question-answering module jointly. The cross-entropy loss between predicted and ground truth answers is used, and the loss is minimized using stochastic gradient descent.

In order to utilize multi-hop facts more effectively, we stacked memory layers and refine question representation q_{k+1} at each layer as sum of output representation at that layer and question representation at previous layer.

$$q_{k+1} = O_{k+1} + q_k.$$

The question representation at layer-1 is obtained using BLSTM. Further, the input and output embeddings are the same across different layers, and at top layer (layer K), answer is predicted as follows.

$$\hat{a} = softmax(f_\theta(O_K + q_{K-1})).$$

Note that we stack three memory layers $(K = 3)$ in our implementation.

The alternatives to memory network is to use BLSTMs to represent knowledge facts. We also use this way of representation as one of our baselines. Here, we sequentially fed multi-hop facts as input to BLSTM by making sure that 1-hop facts always precede 2-hop and so on. This strategy is similar to (Weston, Chopra, and Bordes 2014).

The quantitative results of baselines on KVQA are shown in Table 4. We have shown results in open-world and closed-world settings. Under both these settings, we show results with presence and absence of Wikipedia captions represented as +wikiCap and -wikiCap, respectively. We also evaluate methods on an interesting setting represented as PRP, where our original questions (ORG) are para-phrased in test set. This setting is more realistic as often same questions are asked in many different ways in a real-word setup. We also show results in oracle setting where visual entity linking problem is assumed to be solved. In other words, in

	Method	Oracle	-wikiCap		+wikiCap	
			ORG	PRP	ORG	PRP
Closed World	BLSTM	51.0	47.2	25.0	48.0	27.2
	MemNet	59.2	49.5	32.0	50.2	34.2
Open World	BLSTM	–	16.8	13.1	20.6	14.0
	MemNet	–	36.0	26.2	36.8	30.5

Table 4: Question Answering results on KVQA. Here (i) Oracl: face recognition is solved, (ii) +wikiCap, -wikiCap: wikiCaptions are available and not available, respectively, (iii) ORG: original questions and (iv) PRP: questions are paraphrased. Please refer to the "Visual Entity Linking" section for more details.

Category	ORG	PRP	Category	ORG	PRP
Spatial	48.1	47.2	Multi-rel.	45.2	44.2
1-hope	61.0	60.2	Subtraction	40.5	38.0
Multi-hop	53.2	52.1	Comparison	50.5	49.0
Boolean	75.1	74.0	Counting	49.5	48.9
Intersect.	72.5	71.8	Multi-entity	43.5	43.0

Table 5: VQA results on different categories of questions in paraphrased (PRP) and original version (ORG) of questions in KVQA tested using MemNet. Please see the "VQA over KG" section for more details.

this setting we use ground truth annotation of persons and their order in the image to demonstrate QA results. In all the settings, we assume a fixed size vocabulary and each word token in question, Wikipedia caption, and visual entities are represented as 1-hot vectors.

We observe that BLSTMs are ineffective when number of facts increase and memNet clearly outperforms BLSTM on all settings. Analyzing results more closely, we see that performance goes down by more than 10% when going from closed world to open world. This fall in performance is due to following two reasons: (i) visual entity linking in open-world often makes mistake due to presence of large number of distractors, and (ii) choosing relevant fact for a question given a large number of facts becomes challenging.

As an ablation study, we also report question category-wise results in Table 5. To study the challenges only due to logical reasoning required over KG, we report this result for oracle setting. Recall that in oracle setting we assume persons and their order in image are known to us. Our results suggest that memNet is inadequate in handling spatial, multi-hop, multi-relational, subtraction, counting and multi-entity questions, while it performs well on 1-hop, Boolean and intersection questions. This calls for a need of question-guided approach for this task where sophisticated methods can be designed based on question type. We leave this as one of our future works.

Discussion and Summary

Our results suggest that despite progress in face recognition and question answering, significant further research is needed to achieve high performance on knowledge-aware VQA. The highly-successful off-the-shelf face descriptors, such as Facenet (Schroff, Kalenichenko, and Philbin 2015) falls short in addressing 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.

To sum up, we have drawn attention of the community towards *Knowledge-aware Visual Question Answering* by introducing a challenging dataset – KVQA. We also provided evaluation protocols and baselines for visual named entity linking and question answering over KVQA. The current version of KVQA is limited to persons which is one of the prominent entity types in KG. However, KGs such as Wikidata also contain several other interesting named entity types (such as monuments, events, etc.) which we plan to include in future version of our dataset. We look forward to exciting future research on the theme of bridging images and *world knowledge* inspired by our KVQA dataset and benchmark tasks.

Acknowledgments: Authors would like to thank MHRD, Govt. of India and Intel Corporation for partly supporting this work. Anand Mishra would like to thank Google for supporting conference travel.

References

Acharya, M.; Kafle, K.; and Kanan, C. 2019. TallyQA: Answering complex counting questions. In *AAAI*.

Aditya, S.; Yang, Y.; and Baral, C. 2018. Explicit reasoning over end-to-end neural architectures for visual question answering. In *AAAI*.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: Visual Question Answering. In *ICCV*.

Bharadwaj, S.; Vatsa, M.; and Singh, R. 2014. Aiding face recognition with social context association rule based re-ranking. In *IJCB*.

Cao, Q.; Shen, L.; Xie, W.; Parkhi, O. M.; and Zisserman, A. 2018. Vggface2: A dataset for recognising faces across pose and age. In *FG*.

Fang, Y.; Kuan, K.; Lin, J.; Tan, C.; and Chandrasekhar, V. 2017. Object detection meets knowledge graphs. In *IJCAI*.

G. Narasimhan, M., and Schwing, A. 2018. Straight to the facts: Learning knowledge base retrieval for factual visual question answering. In *ECCV*.

Geman, D.; Geman, S.; Hallonquist, N.; and Younes, L. 2015. Visual turing test for computer vision systems. *Proceedings of the National Academy of Sciences* 112(12):3618–3623.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the V in VQA matter: Elevating

the role of image understanding in Visual Question Answering. In *CVPR*.

Guo, Y.; Zhang, L.; Hu, Y.; He, X.; and Gao, J. 2016. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In *ECCV*.

Hu, P., and Ramanan, D. 2017. Finding tiny faces. In *CVPR*.

Johnson, J.; Hariharan, B.; van der Maaten, L.; Fei-Fei, L.; Zitnick, C. L.; and Girshick, R. 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*.

Kembhavi, A.; Seo, M.; Schwenk, D.; Choi, J.; Farhadi, A.; and Hajishirzi, H. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *CVPR*.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.; Shamma, D. A.; Bernstein, M. S.; and Fei-Fei, L. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123(1):32–73.

Lee, C.-W.; Fang, W.; Yeh, C.-K.; and Wang, Y.-C. F. 2018. Multi-label zero-shot learning with structured knowledge graphs. In *CVPR*.

Lin, D.; Kapoor, A.; Hua, G.; and Baker, S. 2010. Joint people, event, and location recognition in personal photo collections using cross-domain context. In *ECCV*.

Liu, H., and Singh, P. 2004. Conceptnet–a practical commonsense reasoning tool-kit. *BT technology journal* 22(4):211–226.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *ICCV*.

Lu, P.; Ji, L.; Zhang, W.; Duan, N.; Zhou, M.; and Wang, J. 2018. R-VQA: Learning visual relation facts with semantic attention for visual question answering. *arXiv preprint arXiv:1805.09701*.

Malinowski, M., and Fritz, M. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. In *NIPS*.

Marino, K.; Salakhutdinov, R.; and Gupta, A. 2017. The more you know: Using knowledge graphs for image classification. In *CVPR*.

Miller, G. A. 1995. Wordnet: A lexical database for english. *Commun. ACM* 38(11):39–41.

O'Hare, N., and Smeaton, A. F. 2009. Context-aware person identification in personal photo collections. *IEEE Transactions on Multimedia* 11(2):220–228.

Ren, M.; Kiros, R.; and Zemel, R. S. 2015. Image question answering: A visual semantic embedding model and a new dataset. *CoRR* abs/1505.02074.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. FaceNet: A unified embedding for face recognition and clustering. In *CVPR*.

Shen, W.; Wang, J.; and Han, J. 2015. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Trans. Knowl. Data Eng.* 27(2):443–460.

Simonyan, K., and Zisserman, A. 2014. Two-stream convolutional networks for action recognition in videos. In *NIPS*.

Su, Z.; Zhu, C.; Dong, Y.; Cai, D.; Chen, Y.; and Li, J. 2018. Learning visual knowledge memory networks for visual question answering. In *CVPR*.

Tapaswi, M.; Zhu, Y.; Stiefelhagen, R.; Torralba, A.; Urtasun, R.; and Fidler, S. 2016. Movieqa: Understanding stories in movies through question-answering. In *CVPR*.

Toshev, A., and Szegedy, C. 2014. Deeppose: Human pose estimation via deep neural networks. In *CVPR*.

Trott, A.; Xiong, C.; and Socher, R. 2018. Interpretable counting for visual question answering. In *ICLR*.

Vrandecic, D., and Krötzsch, M. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57(10):78–85.

Wang, P.; Wu, Q.; Shen, C.; Dick, A. R.; and van den Hengel, A. 2017. Explicit knowledge-based reasoning for visual question answering. In *IJCAI*.

Wang, P.; Wu, Q.; Shen, C.; Dick, A.; and van den Hengel, A. 2018. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*.

Weston, J.; Chopra, S.; and Bordes, A. 2014. Memory networks. *CoRR* abs/1410.3916.

Zhang, K.; Zhang, Z.; Li, Z.; and Qiao, Y. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* 23(10):1499–1503.

Zhong, Y.; Arandjelović, R.; and Zisserman, A. 2016. Faces in places: Compound query retrieval. In *BMVC*.

Zhu, Y.; Groth, O.; Bernstein, M. S.; and Fei-Fei, L. 2016. Visual7w: Grounded question answering in images. In *CVPR*.