Knowledge Acquisition for Visual Question Answering via Iterative Querying

Yuke Zhu[1]   Joseph J. Lim[2]   Li Fei-Fei[1]

[1]Department of Computer Science, Stanford University

[2]Department of Computer Science, University of Southern California

Abstract

Humans possess an extraordinary ability to learn new skills and new knowledge for problem solving. Such learning ability is also required by an automatic model to deal with arbitrary, open-ended questions in the visual world. We propose a neural-based approach to acquiring task-driven information for visual question answering (VQA). Our model proposes queries to actively acquire relevant information from external auxiliary data. Supporting evidence from either human-curated or automatic sources is encoded and stored into a memory bank. We show that acquiring task-driven evidence effectively improves model performance on both the Visual7W and VQA datasets; moreover, these queries offer certain level of interpretability in our iterative QA model.

1. Introduction

Imagine that you asked your 5-year-old niece "what is the color of the food inside the pan?" to check if the food is ready. Then she saw something like the image in Fig. 1. Unfortunately, this poor kid didn't know what a pan is, and had to ask you "which one is the pan?" before answering that the food is yellow. Just like in this scenario above, our daily interactions often involve asking follow-up questions and collecting "clues" – in order to communicate with others, answer their questions, and serve their needs.

Can machines also ask for and collect "clues" to solve a visual question answering (VQA) task? Most of today's VQA methods make their predictions based on a predefined set of information, typically a mixed representation of the image and the question sentence [3, 9, 23, 39, 41]. As a result, these VQA models have been shown to be "myopic" (tend to fail on novel instances) [1]. Meanwhile, current state-of-the-art VQA models have also indicated that they could benefit from better visually grounded evidence [9, 17]. Hence, we push one step further by enabling a VQA model to ask for and collect "clues" – in particular, visually grounded evidence from human-curated or algorithmically generated data sources.

Figure 1: A human being can understand a scene better by actively asking relevant questions to gain more background information behind the scene, as illustrated in the dialog above. Inspired by this, we propose a dynamic VQA model that can ask queries to get supporting evidence for the task.

Till now, deep learning-based models have dominated standard VQA benchmarks [3, 25, 31, 41]. Among these models, one of the most popular choices is to use CNN to encode images and LSTM to encode words [3, 26, 31]. Furthermore, attention mechanism has been adopted by many top-performing models [9, 23, 41] to achieve better results. Recently Jabri et al. [17] proposed a new alternative model, a two-layer MLP that takes answers as input and makes binary predictions. This simple network has shown highly competitive results in comparison to other more complex architectures. We extend their model with our proposed iterative querying framework to gather and reason about supporting evidence to tackle the VQA tasks.

While our goal is to gather evidence to solve VQA tasks, not all evidence is equally valuable. In fact, most pieces of evidence are irrelevant, and only a few of them are helpful. Therefore, a model has to be selective – ask for and use only relevant information to the task. To this end, we propose a dynamic VQA model that can iteratively ask queries for new evidence and collect relevant evidence from external sources. To be more specific, our model obtains supporting evidence through a series of *queries* from external auxiliary data, called *knowledge sources*. The acquired evidence is encoded and added to a *memory bank*. Then, the model with the newly updated memory can propose another round

of queries, or produce an answer to the target question.

Our experiments show that our model can work well with both human-curated knowledge sources, such as Visual Genome scene graphs [20], and algorithmically generated knowledge sources by the state-of-the-art object detectors [32]. In spite of its simplicity our model achieves new state-of-the-art performance on the Visual7W telling task [41], as well as on par with the top-performing model [9] on the VQA Real Multiple Choice challenge. Another advantage of our model is its interpretability. At every iteration, the model actively seeks new evidence with a textual query. It enables us to examine the model's "rationale" in its iterative process of seeking the final answer.

2. Related Work

VQA Models. Existing VQA models vary from symbolic approaches [25, 38], neural-based approaches [9, 23, 24, 26, 31], to a hybrid scheme of the former two [2]. Attention mechanisms [9, 23, 39, 41] have been shown effective in fusing the multimodal representations of the question words and the images. In addition to these models, some efforts have been spent on better understanding the behavior of existing VQA models [1], as well as evaluating model attention maps against human attention [8]. Jabri et al. [17] proposed a simple alternative model that takes answers as input and performs binary predictions. Their model competes well with other more complex VQA systems. Our work extends their model [17] with a memory bank, achieving better performances on both the Visual7W dataset [41] and the VQA challenge [3].

Interactive Knowledge Acquisition. Knowledge acquisition has been a major interest of AI research for decades. One remarkable pioneer work, dating back to the 1970s, is SHRDLU [36], which provided a dialog system for users to query a computer about the state of a simplified blocks world. Other works have developed interactive interfaces to acquire knowledge from human experts [37], to efficiently label new training samples [35], or to propose the next question in a restricted visual Turing test [10]. Another line of work is never-ending learning, such as NELL [6] and NEIL [7]. However, knowledge harvested in a never-ending loop is often arbitrary and suffers from semantic drift. Several works have investigated a variety of strategies to acquire knowledge from external sources [4, 14, 19, 29, 30, 34]. In contrast to previous work, our work builds a neural-based framework, capable of handling multimodal data. Instead of devising a handcrafted query strategy in previous work, we learn a query strategy in a data-driven fashion.

Memory Networks. A large amount of efforts have devoted to augmenting neural networks with memory. An early prominent innovation is long short-term memory [15], which introduces memory cells to vanilla recurrent neural networks. Recent work [11, 12, 18, 21, 33, 40] focuses on developing different types of external memory representations based on attention mechanism. A work, similar to ours, is dynamic memory network [5], which has an episodic memory module to encode task-dependent information for VQA. However, this work only encodes image features into the episodic memory module, rather than incorporating semantic information. In addition, unlike ours, their model does not learn a strategy to select task-driven evidence based on its relevance to the task.

3. Methods

Many visual questions require open-ended common sense reasoning [3, 10, 16, 41]. A recent study on today's VQA models [1] revealed that most models are "myopic" such that they fail on sufficiently novel concepts. Instead of learning within a closed set, it requires a more flexible and principled model that learns and reasons with an assortment of new information. To the end, our goal is to design a model that proposes **queries** and acquires task-driven evidence (Fig. 2) from knowledge source – specifically, we focus on visually grounded evidence from human-curated or algorithmically generated data sources. We introduce a model that dynamically and constantly learns from external environments in a multi-step fashion. The key challenge here is to learn a querying strategy to gather the most informative evidence for the task.

3.1. Model Overview

Our goal is to iteratively obtain task-driven evidence in order to produce an **answer** to a given visual **question**. This process requires a model that can ask for necessary information from the external sources. Here, we use **query** and **response** as the means of communication between the model and the knowledge sources. Also, the model needs to encode new evidence from a **response** in its own internal representation, which we call **memory**. We define these data types as the following:

- **question**: a natural language question that the model aims to answer about the image. We use a sentence-image pair (q, i) to denote an individual question q on image i;

- **answer**: a natural language answer to the question. We consider the multiple choice tasks, where the goal is to select the correct one from a set of candidate answers;

- **query**: a sentence describing a piece of task-driven information that the model is requesting, e.g., "*How does the man's shirt look?*";

- **response**: a response to a query that contains a piece of evidence from knowledge sources, e.g., "*Striped.*";
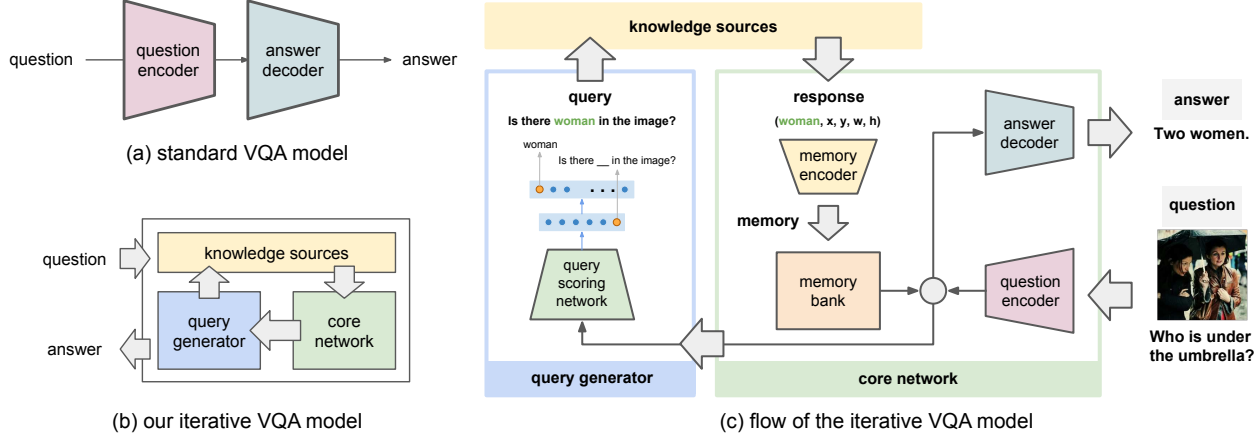
Figure 2: (a) An illustration of a standard VQA model. (b) An overview of our iterative model. (c) Detailed flowchart of our model. The model consists of two major components: core network (green) and query generator (blue). The query generator proposes task-driven queries to fetch evidence from external sources. Acquired knowledge is encoded and stored as memories in the core network for answering a question.

- **memory**: an encoded evidence. Raw evidence is encoded into a vector of memory that the model can store and process.

There are two major challenges in designing an iterative VQA model: 1) proposing the next query at the current model state, and 2) updating the model state with acquired evidence, potentially in various forms from different sources. Our model consists of **core network** (Sec. 3.2) and **query generator** (Sec. 3.3), as shown in Fig. 2(b). The core network handles updating the memory state and generating an answer, while the query generator handles proposing the next query based on the memory state.

3.2. Core Network

The core network is the main component of our model. It takes a question input, predicts an answer, and also maintains its internal memory bank while iteratively obtaining new evidence through querying (see Fig. 3). While the entire core network is jointly trainable in an end-to-end manner, we describe its four sub-networks separately based on their roles:

1. **memory encoders** $f_{\mathcal{K}}$ transform raw evidence e into a memory vector $m = f_{\mathcal{K}}(e)$ that the memory bank can store and process. Raw evidence can be heterogeneous and multimodal. We encode different types of evidence into vectors of the same size;

2. **memory bank** $\mathcal{M}$ stores a collection of memories acquired via iterative querying, where $\mathcal{M} = \{m^{(1)}, m^{(2)}, \ldots, m^{(t)}\}$. The memory bank supports both read/write operations. It can generate a representation of current memory state $\phi_{\mathcal{M}}$ (read). Also, a

new memory can be encoded and added to the memory bank, where $\mathcal{M} := \mathcal{M} \cup \{e^{(t+1)}\}$ (write);

3. **question encoder** $\mathcal{E}_q$ encodes a question-image pair into a vector embedding $v = \mathcal{E}_q(q, i)$;

4. **answer decoder** $\mathcal{G}_a$ takes the question encoding v and the memory state $\phi_{\mathcal{M}}$, and produces an answer $a = \mathcal{G}_a(v, \phi_{\mathcal{M}})$. The question encoder and answer decoder can also be coupled in a single network [17].

Fig. 2 illustrates the interactions between these subnetworks. The formulation of the core network above provides a generic framework, where the design of each subnetwork is modularized. In this work, we demonstrate the effectiveness of our model even without complex network design. We use a simple MLP model [17] as the question encoder and answer decoder. This model is a two-layer MLP, which competes well with state-of-the-art models. It takes as input a concatenation of pretrained image features [13], an average of word embeddings of the question and the answers, and predicts whether an image-question-answer triplet is correct. The memory encoder transforms raw evidence from external knowledge sources into fixed-dimensional memory vectors. We represent each memory as a 300-dimensional averaged word2vec embedding [27]. We provide more details of memory encoders in the supplementary material. To retain the simplicity of our model, we use a stack as our memory bank. It keeps the encoded memory vectors, and updates itself by adding a new memory to the stack. We compute the memory state by summing the memory vectors, normalized by ℓ_2-norm, in the memory bank. We concatenate this memory state vector with the image-question-answer triplet as input to our MLP model.

Q: What object can you see?
R: (woman, 204, 35, 54, 103)

Q: What is interacting with woman?
R: (woman, wearing, wedding gown)

When was this picture taken?
A. During a wedding.
B. During a bar mitzvah.
C. During a funeral.
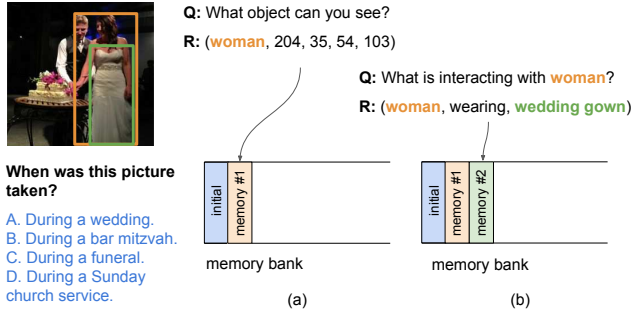D. During a Sunday church service.

(a)   (b)

Figure 3: An example iterative query process. At each time, the model proposes a task-driven query (Q) to request useful evidence from knowledge sources. The response (R) is encoded into a memory and added to the memory bank.

3.3. Query Generator

The query generator bridges our model with the external knowledge sources. It proposes queries based on the memory state to obtain best relevant evidence. While the most straightforward strategy would be to paraphrase the target question as a query to an omniscient source, we don't have such oracle in practice. Hence, it is essential to define useful query types for communication and to devise a good query strategy for effectiveness.

An error analysis in previous work [17] indicates that a lack of visual grounding, i.e., facts about the objects in an image, is a key problem in current VQA systems. Such visual grounding would help resolving the underlying uncertainty from noisy vision models. Hence, we define four query types that the model can use to request visually grounded evidence. In Table 1, the bold words in the query templates are the free arguments to construct the final queries. Such evidence in the responses is sometimes referred to as *episodic memories* [21], as they are grounded on a specific image. These responses can be harvested from either human annotation or a pretrained predictive model.

Table 1: Query Types and Response Formats

Query types and templates	Response formats
What object can you see?	(**object**, x, y, w, h)
Is there **object** in the image?	(**object**, x, y, w, h)
How does **object** look?	(**object**, **attribute**)
What is interacting with **object1**?	(**object2**, **relation**)

Now, we need a strategy to generate the best query to ask at the current memory state. Reinforcement learning (RL) approaches are commonly used to learn such a querying policy. However, we found that standard deep RL methods such as DQN [28] have convergence issues in our problem setting with a large discrete action space. To ad-

dress this limitation, we use a tree expansion method with a greedy scoring function instead. We use supervised learning method to train a query scoring network, which evaluates query candidates at the current state.

Our query scoring network is an MLP model, similar to the core network, followed by two-level hierarchical softmax for the query types and the query objects correspondingly (see Fig. 4(a)). It takes an image-question-memory triplet as input; however in contrast to the core network, it does not take answer vectors as input. As we don't have ground-truth labels of the optimal queries at each step, we automatically generate the training samples by Monte-Carlo rollouts. Fig. 4(b) demonstrates a rollout procedure of the query tree expansion method. Each node in the tree represents a query candidate. At each step, we maintain a set of nouns that have been seen in question and responses, and branch out queries from this set. The noun set is initialized by all the noun entities in the question. This set constrains the width of the search tree, making computation tractable. During test, the query scoring network computes a score for each terminal node. The model proposes the next query with the highest score.

3.4. Learning

As mentioned in Sec. 3.2, the core network, therefore, can be trained end-to-end. However, at each step, the query generator makes a hard decision on which query to propose, introducing a non-differentiable operation, yet there exists interdependence between the core network and the query generator. Thus, we devise an EM-style training procedure, where we freeze the core network while training the query scoring network, and vice versa (see Algorithm 1).

We bootstrap with a uniformly random strategy as the seed query generator, as we initially don't have a trained query scoring network. The initial core network is trained with random rollouts using backpropagation. In subsequent iterations, the core network is trained with rollouts generated from a trained query generator (i.e., tree expansion + query scoring network) from previous step. Freezing the core network, we then train the query scoring network.

We train the query scoring network with the image-question-memory triplets as input. The training set is automatically generated by the core network on Monte-Carlo rollouts, as depicted in Fig. 4(c). In each rollout, we add a pair of input and label (i.e., query type and query object) to the training set if the newly added memory flipped previously incorrect predictions to the correct answers.

3.5. Implementation Details

We follow the same network setup and the same hyperparameters as [17]. Both the core network and the query scoring network have 8,192 hidden units. We use dropout (0.5) after the first layer, and ReLU as the non-linearity.

Algorithm 1 Training Procedure for Iterative QA Model

1: **procedure**
2:     Generate random query rollouts $R^{(0)}$
3:     Train initial core network $\mathcal{C}^{(0)}$ with rollout $R^{(0)}$
4:     Generate training samples $S^{(0)}$ for query scoring network with $\mathcal{C}^{(0)}$
5:     Train initial query scoring network $\mathcal{G}^{(0)}$ with $S^{(0)}$
6:     **for** $t = 1, \ldots, N$ **do**   ▷ Iterate N times
7:         Generate query rollouts $R^{(t)}$ with query scoring network $\mathcal{G}^{(t-1)}$
8:         Finetune core network $\mathcal{C}^{(t)}$ from $\mathcal{C}^{(t-1)}$ with rollout $R^{(t)}$
9:         Generate training samples $S^{(t)}$ for query scoring network from $\mathcal{C}^{(t)}$
10:       Finetune query scoring network $\mathcal{G}^{(t)}$ from $\mathcal{G}^{(t-1)}$ with $S^{(t)}$
11:     **end for**
12:     **return** $\{\mathcal{G}^{(N)}, \mathcal{C}^{(N)}\}$
13: **end procedure**

(a) Query scoring network   (b) Query search tree expansion   (c) Training sample generation
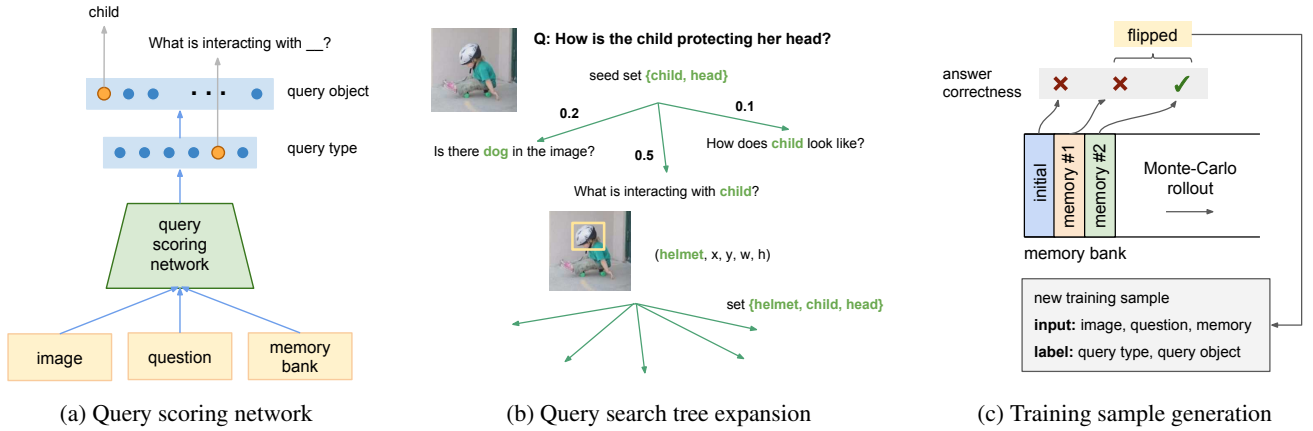
Figure 4: (a) Query scoring network. The network uses hierarchical softmax to evaluate each query given the current memory state and the question; (b) Query search tree expansion. The model starts with a seed set of nouns, which is used to generate queries; new nouns from query responses are added to the set and used to expand the query tree further; (c) Generating training samples for query scoring network. We perform Monte-Carlo rollouts on the query search tree, and use feedback from the core network as the labels.

Both networks are trained using SGD with momentum and a base learning rate of 0.01. We perform Monte-Carlo rollouts by the query generator using an ϵ-greedy strategy (Line 7 in Algorithm 1), where ϵ is annealed from 1.0 to 0.1 as the iterative training procedure ($N = 5$) proceeds.

4. Experiments

Our main goal, throughout experiments, is to examine how the acquired evidence from the iterative QA model impact performances of answering questions on images. We aim to investigate two major aspects of our model: 1) impacts of querying strategies for task-driven knowledge acquisition, and 2) contributions and limitations of different knowledge sources to answering questions on images. We first report quantitative results in Sec. 4.2 and then perform detailed analysis in Sec. 4.3.

4.1. Experiment Setups

Datasets. Our experiments are conducted on the Visual7W telling task [41] and the VQA Real Multiple Choice challenge [3]. The Visual7W telling dataset includes 69,817 questions for training, 28,020 for validation, and 42,031 for testing. The performance is measured by the percentage of questions that are correctly answered. The VQA Real Multiple Choice challenge has 248,349 questions for training, 121,512 for validation, and 244,302 for testing. The performance is reported by an evaluation metric proposed by [3].

Knowledge Sources. As indicated by the error analysis in Jabri et al. [17], the key limitation of today's VQA models is a lack of visual grounding of objects and concepts. We thus design our query types and responses in Table 1 to acquire visually grounding evidence from the knowledge sources. Such knowledge can be obtained by human an-

notation, or automatically generated by pretrained models. Based on the availability of knowledge sources, we evaluate our model with human-curated knowledge sources [20] for the Visual7W dataset, and with automatic knowledge sources [32] for the VQA challenge.

The Visual7W dataset is collected on a subset of images from Visual Genome [20]. Visual Genome offers a structured image representation called *scene graphs*, a graph structure of objects, attributes of objects, and their pairwise relationships. These scene graphs are manually annotated by AMT workers. We use these scene graphs as the knowledge source for Visual7W. We use these ground-truth annotations of the most frequent 150 objects, 50 attributes, and 20 relations, as the responses to the queries.

The VQA dataset is collected on images from the COCO dataset [22]. In the absence of scene graph annotations, we run the state-of-the-art object detector [32], trained with 80 object classes, to predict objects on the images. These automatic predictions are used as the responses to the first two types of queries in Table 1. We discard detections with low scores by a cutoff of 0.5. We omit the other query types due to the lack of a reliable scene graph generation model.

4.2. Quantitative Experiments

Our model is built upon a simple MLP model [17]. The main novelty of our model is to augment this model with a memory bank, where task-driven memories are actively obtained by iterative queries. The iterative querying model generates up to three queries before producing the final answer. We examine the performance of our query generator with three querying strategies:

- **all knowledge** populates the memory bank with the entire knowledge source, i.e., the entire scene graph or all detected objects of the image;

- **uniform sampling** randomly proposes queries without using the query generator;

- **query generator** selects queries based on the trained query generator introduced in Sec. 3.4.

We report the performance on the Visual7W telling test split in Table 2 and on VQA Real Multiple-Choice test-dev and test-standard in Table 3. We compare our model with the state-of-the-art VQA models to date. The current state-of-the-art results on Visual7W is reported by Jabri et al. [17]. The winning model on VQA Real Multiple-Choice challenge is Fukui et al. [9]. For Visual7W, we train our model on the training split. For VQA, we train our model on both the *train* set and *train+val* set. We follow the setup of previous work [9, 23], and report the final test-standard performance with our model trained on the *train+val* set.

Table 2 and Table 3 illustrate the effectiveness of our iterative querying model. It achieves new state-of-the-art results in Visual7W and on par with the best single model (MCB + Att. + GloVe [9]) on VQA.[1] Comparing to the two querying strategy baselines, our query generator learns to query selectively, offering the best performances. Our simple MLP model augmented by the memory bank show competitive results compared to existing models [9, 23] that have a much complex design. Besides its simplicity, the iterative queries offer us a chance to interpret the model's "rationale". Fig. 5 shows some qualitative question examples from both datasets. We notice how the answer predictions are changed as the querying process goes when new evidence is being acquired. The model can often correct its previously false predictions when a relevant memory is acquired and added to the memory bank. However, in some cases, a digressive piece of evidence can conversely misguide the model. For instance, a false detection of train in the last example of Fig. 5 causes the model to over-count.

4.3. Model Analysis

Despite of comparable performance gains on Visual7W and VQA, they have used two distinct types of knowledge sources respectively. The Visual7W model uses ground-truth scene graph annotations (including testing phase) from Visual Genome [20], which is costly and tedious to collect. In contrast, the VQA model uses predictions from faster R-CNN detectors [32], which is cheap and efficient to obtain.

One intuitive explanation of the modest performance gain by using ground-truth scene graphs is the sparsity and ambiguity of human annotation. Scene graphs are sparse (e.g., about 20 objects per image) and open-vocabulary (e.g, "kid" vs "boy"). Consequently, only 28% objects in answers can be mapped to scene graphs. Furthermore, based on the grounding annotations of Visual7W, only 43% answers mention at least one object. Hence, a naïve keyword matching baseline with random tie breaker would yield a poor accuracy of 35.7%, only 10% above chance. In contrast, the automatic knowledge sources generated by predictive models do not suffer from the sparsity and ambiguity of human annotation. A perfect object detector would be able to find every object instances within its predefined vocabulary. However, in reality its value is undermined by the imperfect performance of these models. For instance, the faster R-CNN detector that we used has 42.7% mAP@.5 on COCO test-dev. Thus, we observe that the trade-off of the hand-crafted and automatic knowledge sources from different perspectives. An ideal knowledge source would combine the strengths of both types, and provide a precise and

---

[1] The best number reported on the VQA challenge to date is 0.701 on *test-standard* by Fukui et al. [9]. However, this model is an ensemble of 7 MCB models that trained with additional QA pairs from Visual Genome [20], which is only 1.2% better than our results.

Table 2: Model Performance on the Visual7W test split

method	what	where	when	who	why	how	overall
LSTM-Attention [41]	0.515	0.570	0.750	0.595	0.555	0.498	0.543
MCB [9]	0.603	0.704	0.795	0.692	0.582	0.511	0.622
MLP [17]	0.628	0.735	0.797	0.709	0.623	0.538	0.648
MLP + all knowledge	0.633	0.741	0.806	0.752	0.644	0.540	0.658
MLP + uniform sampling	0.624	0.740	0.805	0.762	0.629	0.537	0.653
MLP + query generator	0.651	0.778	0.807	0.814	0.653	0.541	0.679

Table 3: Model Performance on the VQA test-dev and test-standard

method	test-dev				test-standard			
	yes/no	number	other	all	yes/no	number	other	all
Two-layer LSTM [3]	-	-	-	0.627	0.806	0.377	0.536	0.631
Co-Attention [23]	-	-	-	0.658	0.800	0.395	0.599	0.661
MCB + Att. + GloVe [9]	-	-	-	0.691	-	-	-	-
MCB Ensemble + Genome [9]	-	-	-	0.702	0.833	0.410	0.652	0.701
MLP [17]	0.787	0.402	0.608	0.659	-	-	-	-
MLP + all knowledge	0.787	0.405	0.625	0.668	-	-	-	-
MLP + uniform sampling	0.788	0.404	0.622	0.666	-	-	-	-
MLP + query generator	0.803	0.395	0.626	0.674	-	-	-	-
MLP + query generator (train+val)	0.814	0.421	0.646	0.691	0.814	0.417	0.642	0.689

complete coverage of the visual concepts in an image.

We further analyze the limits of our visually grounded evidence in improving VQA performance. We hypothesize that the VQA tasks cannot be completely reduced to a visual grounding problem, as some of the questions involve common sense reasoning about novel concepts [1]. To test our hypothesis, we conduct a human study where we asked 5 human subjects to answer questions from both datasets. These subjects are shown the entire knowledge sources for each image, without seeing the images. They are asked to select the best multiple choices of each question given the knowledge sources. We randomly sample 500 questions from each question type (six types in Visual7W and three types in VQA). The human accuracy is reported as the majority vote among the subjects. The results (Q + KS) is reported in Table 4. We compare with previous records [3, 41] of majority human performances when subjects answered questions without images (Q) and with images (Q + I).

Table 4: Ablation Study of Human Performance

	Q	Q + I	Q + KS
Visual7W	0.353	0.957	0.522
VQA	-	0.879	0.476

Not surprisingly, our knowledge sources greatly improved human performance without showing images. Yet they still do not compensate the absence of images. We observe more than 40% performance gap between Q + KS and Q + I on both datasets. This validates our hypothesis that our knowledge sources cannot obsolete the rich visual

content of images for the VQA tasks. We provide human performance by question type in the supplementary material. We observe that our model and human subjects exhibit different patterns on different question types. Our model offers the most performance gain (10.5%) over the baseline on *who* questions, as persons are among the most common object categories in scene graphs [20]. However, human subjects have the largest improvement (25.4%) on *where* questions by seeing the scene graph, as they can often infer the scene category based on the objects (e.g., inferring "baseball field" after seeing mitt and bat). That implies, humans can take advantage of the knowledge sources in a more sophisticated way, e.g., utilizing common sense to jointly reason about a large variety of concepts.

In this work, we intentionally retain the simplicity of the model design while showing its effectiveness. Our simple MLP model requires the presence of multiple choices as input. As our model is modularized, each component can be independently substituted by a more complex system (e.g., a sequence generator as the answer decoder for open-ended VQA tasks). Our analysis sheds light on two possible directions to improve our model: 1) to augment our knowledge sources to increase the variety and coverage of visual concepts; and 2) to explore better approaches to encoding memory with common sense. Furthermore, our proposed iterative querying model can be viewed as a generic framework to acquire task-driven information, and is easily plugged into other types of visual tasks by appending the memory bank to a predictive model. A future direction would be to explore the potential of our model in other tasks.

baseline | query #1 | query #2 | query #3

Visual7W

What color drink is in the nearest glass?
Clear.
Red.
Brown.
Blue.

Is there glass in the image?
(**glass**, x, y, w, h)

Is there glass in the image?
(**glass**, x, y, w, h)

What is interacting with glass?
(**water**, **of**, **glass**)

→ time

Prediction: **Brown.** | **Blue.** | **Blue.** | **Clear.**

Visual7W

Where is the plate?
On the floor.
In the dishwasher.
On the table.
In the cabinet.

Is there plate in the image?
(**plate**, x, y, w, h)

What new object can you see?
(**floor**, x, y, w, h)

How does plate look?
(**plate**, green)

→ time

Prediction: **In the cabinet.** | **On the table.** | **On the floor.** | **On the floor.**

Visual7W

Who is holding the orange?
A man.
A girl.
A boy.
A woman.

What object can you see?
(**man**, x, y, w, h)

What object can you see?
(**woman**, x, y, w, h)

Is there orange in the image?
(**orange**, x, y, w, h)

→ time

Prediction: **A man.** | **A man.** | **A woman.** | **A woman.**

VQA

What would happen if the driver pressed the accelerator?
Backwards
Dead sheep
Movement

Is there chair in the image?
(**chair**, x, y, w, h)
Note: false detection

Is there sheep in the image?
(**sheep**, x, y, w, h)

Is there sheep in the image?
(**sheep**, x, y, w, h)

→ time

Prediction: **Backwards.** | **Movement.** | **Dead sheep.** | **Dead sheep.**

VQA

Is there anything in this scene used to put out fires?
Yes
No
...

Is there bicycle in the image?
(**bicycle**, x, y, w, h)

Is there person in the image?
(**person**, x, y, w, h)

Is there fire hydrant in the image?
(**fire hydrant**, x, y, w, h)

→ time

Prediction: **No.** | **No.** | **No.** | **Yes.**

VQA

How many trains are there?
1
0
...

What object can you see?
(**clock**, x, y, w, h)

Is there train in the image?
(**train**, x, y, w, h)
Note: false detection

What object can you see?
(**clock**, x, y, w, h)

→ time

Prediction: **0.** | **0.** | **1.** | **1.**

Figure 5: Qualitative results of our final model. We show the model's predictions with no query (i.e., MLP baseline [17]), one, two and three queries. The left shows the question, the image, and a subset of multiple choices, followed by three queries proposed by our model. Answers below the arrows are predictions at each time step, where green shows correct predictions, and red incorrect. We showcase examples where our model switches to correct/incorrect answers in the querying process.

5. Conclusion

We propose a new scheme to tackle the task of visual QA via iterative knowledge acquisition. Our model actively acquires new evidence from external sources via task-driven querying. Our experiments have shown that the model manages to leverage newly obtained evidence and significantly boosts the performance of answering visual questions. Our model is a preliminary attempt for a system that learns to interact with external environments for prolonged, continuous learning. Future directions include exploration of better ways to represent common sense, to harvest information from less curated knowledge sources, and to generalize our model to other problem domains.

References

[1] A. Agrawal, D. Batra, and D. Parikh. Analyzing the behavior of visual question answering models. *EMNLP*, 2016. 1, 2, 7

[2] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Neural module networks. In *CVPR*, 2016. 2

[3] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual question answering. *ICCV*, 2015. 1, 2, 5, 7

[4] S. Branson, C. Wah, F. Schroff, B. Babenko, P. Welinder, P. Perona, and S. Belongie. Visual recognition with humans in the loop. In *ECCV*, 2010. 2

[5] R. S. Caiming Xiong, Stephen Merity. Dynamic memory networks for visual and textual question answering. *ICML*, 2016. 2

[6] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka, and T. M. Mitchell. Toward an architecture for never-ending language learning. In *AAAI*, 2010. 2

[7] X. Chen, A. Shrivastava, and A. Gupta. NEIL: Extracting visual knowledge from web data. *ICCV*, 2013. 2

[8] A. Das, H. Agrawal, C. L. Zitnick, D. Parikh, and D. Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *EMNLP*, 2016. 2

[9] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *EMNLP*, 2016. 1, 2, 6, 7

[10] D. Geman, S. Geman, N. Hallonquist, and L. Younes. Visual turing test for computer vision systems. *Proceedings of the National Academy of Sciences*, 112(12):3618–3623, 2015. 2

[11] A. Graves, G. Wayne, and I. Danihelka. Neural turing machines. *CoRR*, abs/1410.5401, 2014. 2

[12] E. Grefenstette, K. M. Hermann, M. Suleyman, and P. Blunsom. Learning to transduce with unbounded memory. In *NIPS*, 2015. 2

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CVPR*, 2016. 3

[14] B. Hixon, P. Clark, and H. Hajishirzi. Learning knowledge graphs for question answering through conversational dialog. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015. 2

[15] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2

[16] T.-H. K. Huang, F. Ferraro, N. Mostafazadeh, I. Misra, A. Agrawal, J. Devlin, R. Girshick, X. He, P. Kohli, D. Batra, C. L. Zitnick, D. Parikh, L. Vanderwende, M. Galley, and M. Mitchell. Visual storytelling. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016. 2

[17] A. Jabri, A. Joulin, and L. van der Maaten. Revisiting visual question answering baselines. In *ECCV*, 2016. 1, 2, 3, 4, 5, 6, 7, 8

[18] A. Joulin and T. Mikolov. Inferring algorithmic patterns with stack-augmented recurrent nets. In *NIPS*, 2015. 2

[19] J. Kim and Y. Gil. Incorporating tutoring principles into interactive knowledge acquisition. *International Journal of Man-Machine Studies*, 65(10):852–872, 2007. 2

[20] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalanditis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017. 2, 6, 7

[21] A. Kumar, O. Irsoy, J. Su, J. Bradbury, R. English, B. Pierce, P. Ondruska, I. Gulrajani, and R. Socher. Ask me anything: Dynamic memory networks for natural language processing. *ICML*, 2016. 2, 4

[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 6

[23] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. *NIPS*, 2016. 1, 2, 6, 7

[24] L. Ma, Z. Lu, and H. Li. Learning to answer questions from image using convolutional neural network. *AAAI*, 2016. 2

[25] M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *NIPS*, 2014. 1, 2

[26] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. *ICCV*, 2015. 1, 2

[27] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013. 3

[28] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015. 4

[29] K. Narasimhan, A. Yala, and R. Barzilay. Improving information extraction by acquiring external evidence with reinforcement learning. *EMNLP*, 2016. 2

[30] A. Pappu and A. I. Rudnicky. Knowledge acquisition strategies for goal-oriented dialog systems. In *15th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 194, 2014. 2

[31] M. Ren, R. Kiros, and R. Zemel. Exploring models and data for image question answering. *NIPS*, 2015. 1, 2

[32] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 2, 6

[33] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus. End-to-end memory networks. In *NIPS*, 2015. 2

[34] P. Welinder, S. Branson, S. Belongie, and P. Perona. The multidimensional wisdom of crowds. In *NIPS*, 2010. 2

[35] J. D. Williams, N. B. Niraula, P. Dasigi, A. Lakshmiratan, C. G. J. Suarez, M. Reddy, and G. Zweig. Rapidly scaling

dialog systems with interactive learning. In *Natural Language Dialog Systems and Intelligent Assistants*, pages 1–13. Springer, 2015. 2

[36] T. Winograd. Procedures as a representation for data in a computer program for understanding natural language. Technical report, MIT AI Technical Report 235, 1971. 2

[37] M. Witbrock, D. Baxter, J. Curtis, D. Schneider, R. Kahlert, P. Miraglia, P. Wagner, K. Panton, G. Matthews, and A. Vizedom. An interactive dialogue system for knowledge acquisition in cyc. *IJCAI Workshop on Mixed-Initiative Intelligent Systems*, 2003. 2

[38] Q. Wu, P. Wang, C. Shen, A. v. d. Hengel, and A. Dick. Ask me anything: Free-form visual question answering based on knowledge from external sources. *ICML*, 2016. 2

[39] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. *CVPR*, 2016. 1, 2

[40] W. Zaremba and I. Sutskever. Reinforcement learning neural turing machines. *arXiv preprint arXiv:1505.00521*, 2015. 2

[41] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7W: Grounded Question Answering in Images. In *CVPR*, 2016. 1, 2, 5, 7