© 2019 Medhini Gulganjalli Narasimhan

VISUAL QUESTION ANSWERING USING EXTERNAL KNOWLEDGE

BY

MEDHINI GULGANJALLI NARASIMHAN

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2019

Urbana, Illinois

Advisers:

Professor Alexander Schwing
Professor Svetlana Lazebnik

ABSTRACT

Accurately answering a question about a given image requires combining observations with general knowledge. While this is effortless for humans, reasoning with general knowledge remains an algorithmic challenge. To advance research in this direction, a novel 'fact-based' visual question answering (FVQA) task has been introduced recently along with a large set of curated facts which link two entities, *i.e.*, two possible answers, via a relation. Given a question-image pair, keyword matching techniques have been employed to successively reduce the large set of facts and were shown to yield compelling results despite being vulnerable to misconceptions due to synonyms and homographs.

To overcome these shortcomings, we introduce two new approaches in this work. We develop a learning-based approach which goes straight to the facts via a learned embedding space. We demonstrate state-of-the-art results on the challenging recently introduced fact-based visual question answering dataset, outperforming competing methods by more than 5%. Upon further analysis, we observe that a successive process which considers one fact at a time to form a local decision is sub-optimal. To counter this, in our second approach we develop an entity graph and use a graph convolutional network to 'reason' about the correct answer by jointly considering all entities. We show on the FVQA dataset that this leads to an improvement in accuracy of around 7% compared to the state-of-the-art.

To my Mom (Latha), Grandpa (Iyengar) and Dog (Nano).

ACKNOWLEDGMENTS

I owe this thesis to my advisors, Professor Alexander Schwing and Professor Svetlana Lazebnik. I'd like to thank them for all the time they invested in me, teaching me how to do good research and helping me build my career. Their support and encouragement has played a vital role in my growth over the past two years. I could only dream of doing a Ph.D. before I set out on my Master's, and today this is a reality thanks to their help.

I'd like to thank my Grandfather and Mom for helping me through good and bad and for their unconditional love.

I received a great deal of help and support from my peers in the vision lab. I'd like to thank Arun Mallya for the insightful conversations that sparked research ideas and for the much needed help when I first started working on research. I will cherish the fun times spent with my lab mates, who helped me let off steam when work got too stressful.

I would like to thank NVIDIA for providing the GPUs used for this research. This material is based upon work supported in part by the National Science Foundation under Grant No. 1718221 and Grant No. 1563727, Samsung, 3M, IBM-ILLINOIS Center for Cognitive Computing Systems Research (C3SR), Amazon Research Award, and AWS Machine Learning Research Award.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

When answering questions about images, we easily combine the visualized situation with general knowledge that is available to us. However, for algorithms, an effortless combination of general knowledge with observations remains challenging, despite significant work which aims to leverage these mechanisms for autonomous agents and virtual assistants. If successful, participation of agents and virtual assistants in our day to day endeavors seems within reach.

In recent years, a significant amount of research has investigated algorithms for visual question answering (VQA) [1–6], visual question generation (VQG) [7–10], and visual dialog [11–13], paving the way to autonomy for artificial agents operating in the real world. Images and questions in these datasets cover a wide range of perceptual abilities such as counting, object recognition, object localization, and even logical reasoning. However, for many of these datasets the questions can be answered solely based on the visualized content, *i.e.*, no general knowledge is required. Therefore, numerous approaches address VQA, VQG and dialog tasks by extracting visual cues using deep network architectures [1–5,14–29], while general knowledge remains unavailable.

To bridge this discrepancy between human behavior and present day algorithmic design, Wang *et al*. [30] introduced a novel 'fact-based' VQA (FVQA) task, and an accompanying dataset containing images, questions with corresponding answers and a knowledge base (KB) of facts extracted from three different sources: WebChild [31], DBPedia [32] and ConceptNet [33]. Unlike classical VQA datasets, a question in the FVQA dataset is answered by a collective analysis of the information in the image and the KB of facts. Each question is mapped to a single supporting fact which contains the answer to the question. Thus, answering a question requires analyzing the image and choosing the right supporting fact, for which Wang *et al*. [30] propose a keyword-matching technique. The method presented in FVQA [30] produces a query as an output of an LSTM which is fed the question as an input. Facts in the knowledge base are filtered on the basis of visual concepts such as objects, scenes, and actions extracted from the input image. The predicted query is then applied on the filtered database, resulting in a set of retrieved facts. A matching score is subsequently computed between the retrieved facts and the question to determine the most relevant fact. The most correct fact forms the basis of the answer for the question. This approach suffers when the question doesn't focus on the most obvious visual concept and when there are synonyms and homographs. Moreover, special information about the visual concept type and the answer source make it hard to generalize their approach to other datasets.

To address the shortcomings of the previous work [30], we introduce two new approaches

Question: Which object in the image can be used to eat with?
Relation: UsedFor
Associated Fact: (Fork, UsedFor, Eat)
Answer Source: Image
Answer: Fork

Question: What do the animals in the image eat?
Relation: RelatedTo
Associated Fact: (Sheep, RelatedTo, Grass Eater)
Answer Source: Knowledge Base
Answer: Grass

Question: Which equipment in this image is used to hit baseball?
Relation: CapableOf
Associated Fact: (Baseball bat, CapableOf, Hit a baseball)
Answer Source: Image
Answer: Baseball bat

Figure 1.1: Results of our graph convolutional net based approach on the recently introduced FVQA dataset.

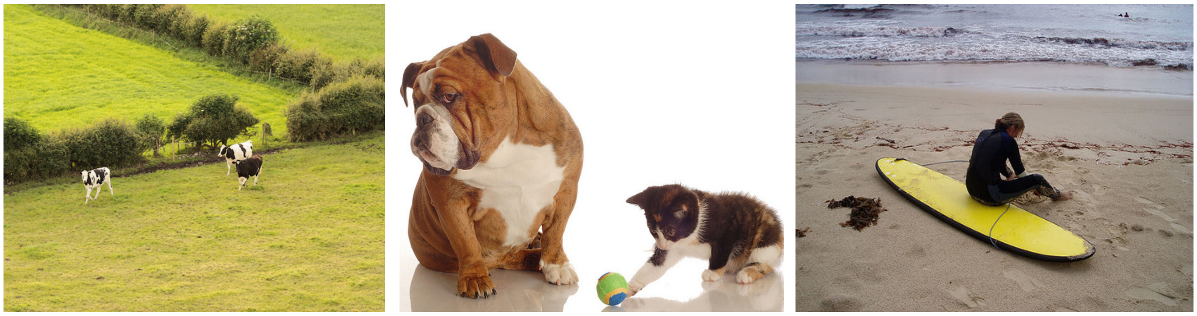
which are described in the following sections.

1.1 LEARNING KNOWLEDGE BASE RETRIEVAL

In [34], we develop a learning based retrieval method. More specifically, our approach learns a parametric mapping of facts and question-image pairs to an embedding space. To answer a question, we use the fact that is most aligned with the provided question-image pair. As illustrated in Fig. 1.1, our approach is able to accurately answer both more visual questions as well as more fact based questions. For instance, given the image illustrated on the left hand side along with the question, "Which object in the image can be used to eat with?", we are able to predict the correct answer, "fork." Similarly, the proposed approach is able to predict the correct answer for the other two examples. Quantitatively we demonstrate the efficacy of the proposed approach on the recently introduced FVQA dataset, outperforming state-of-the-art by more than 5% on the top-1 accuracy metric.

1.2 REASONING WITH GRAPH CONVOLUTION NETWORKS

In [35], our main motivation is to develop a technique which uses the information from multiple facts before arriving at an answer. This technique relies less on retrieving the single 'correct' fact needed to answer a question. To this end, we develop a model which 'thinks out of the box,' *i.e.*, it 'reasons' about the right answer by taking into account a list of facts via a Graph Convolution Network (GCN) [36]. The GCN enables *joint* selection of the answer from a list of candidate answers, which sets our approach apart from the previous methods



Question: What is the area in the image used for?
Relation: UsedFor
Visual Concept: Field
Fact: (Field, UsedFor, Grazing Animals)
Answer: Grazing Animals

Question: Which object in the image is more similar to a tiger?
Relation: RelatedTo
Visual Concept: Cat
Fact: (Cat, RelatedTo, Tiger)
Answer: Cat

Question: What can be found on the ground in this image?
Relation: AtLocation
Visual Concept: Beach
Fact: (Sand, AtLocation, Beach)
Answer: Sand

Figure 1.2: Results of our graph convolutional net based approach on the recently introduced FVQA dataset.

that assess one fact at a time. Moreover, we select a list of supporting facts in the KB by ranking GloVe embeddings. This handles challenges due to synonyms and homographs and also works well with questions that don't focus on the main object.

We demonstrate the proposed algorithm on the FVQA dataset [30], outperforming the state of the art by around 7%. Fig. 1.2 shows results obtained by our model. Unlike the models proposed in [30], our method does not require any information about the ground truth fact (visual concept type and answer source). In contrast to our approach in [34], which focuses on learning a joint image-question-fact embedding for retrieving the right fact, this latest work uses a simpler method for retrieving multiple candidate facts (while still ensuring that the recall of the ground truth fact is high), followed by a novel GCN inference step that collectively assesses all the relevant facts before arriving at an answer. Using an ablation analysis we find improvements due to the GCN component, which exploits the graphical structure of the knowledge base and allows for sharing of information between possible answers, thus improving the explainability of our model.

CHAPTER 2: RELATED WORK

We develop two frameworks for visual question answering that benefit from a rich knowledge base. In the following, we first review classical visual question answering tasks before discussing visual question answering methods that take advantage of knowledge bases. Our second approach [35], is based on graph convolutional nets which benefits from general knowledge encoded in the form of a knowledge base. We therefore briefly review existing work in graph convolutional networks as well.

2.1 VISUAL QUESTION ANSWERING

Recently, there has been significant progress in creating large VQA datasets [1,14,16,37–39] and deep network models which correctly answer a question about an image. The initial VQA models [1–3, 5, 15, 16, 18–28, 38, 40] combined the LSTM encoding of the question and the CNN encoding of the image using a deep network which finally predicted the answer. Results can be improved with attention-based multi-modal networks [2, 3, 6, 18–20, 23, 24] and dynamic memory networks [17, 41]. All of these methods were tested on standard VQA datasets where the questions can solely be answered by observing the image. No out of the box thinking was required. For example, given an image of a cat, and the question, "Can the animal in the image be domesticated?," we want our method to combine features from the image with common sense knowledge (a cat can be domesticated). This calls for the development of a model which leverages external knowledge.

2.2 FACT-BASED VISUAL QUESTION ANSWERING

Recent research in using external knowledge for natural language comprehension led to the development of semantic parsing [34, 42–53] and information retrieval [54–60] methods. However, knowledge based visual question answering is fairly new. Notable examples in this direction are works by Zhu *et al*. [61], Wu *et al*. [62], Wang *et al*. [63], Narasimhan *et al*. [64], Krishnamurthy and Kollar [65], and our previous work, Narasimhan and Schwing [34].

Ask Me Anything (AMA) by Wu *et al*. [62], AHAB by Wang *et al*. [63], and FVQA by Wang *et al*. [30] are closely related to our work. In AMA, attribute information extracted from the image is used to query the external knowledge base DBpedia [32], to retrieve paragraphs which are summarized to form a knowledge vector. The knowledge vector is combined with the attribute vector and multiple captions generated for the image, before being passed as input to an LSTM which predicts the answer. The main drawback of AMA is that it does not perform any explicit reasoning and ignores the possible structure in the KB. To address this, AHAB and FVQA attempt to perform explicit reasoning. In AHAB, the question is converted to a database query via a multistep process, and the response to the query is processed to obtain the final answer. FVQA also learns a mapping from questions to database queries through classifying questions into categories and extracting parts from the question deemed to be important. A matching score is computed between the facts retrieved from the database and the question, to determine the most relevant fact which forms the basis of the answer for the question. Both these methods use databases with a particular structure: facts are represented as tuples, for example, (*Apple*, *IsA*, *Fruit*), and (*Cheetah*, *FasterThan*, *Lion*).

The present work follows up on our earlier method, *Straight to the Facts* (STTF) [34]. STTF uses object, scene, and action predictors to represent an image and an LSTM to represent a question and combines the two using a deep network. The facts are scored based on the cosine similarity of the image-question embedding and fact embedding. The answer is extracted from the highest scoring fact.

We evaluate our method on the dataset released as part of the FVQA work, referred to as the FVQA dataset [30], which is a subset of three structured databases – DBpedia [32], ConceptNet [33], and WebChild [31].

2.3 GRAPH CONVOLUTIONAL NETWORKS

Kipf and Welling [36] introduced Graph Convolutional Networks (GCN) to extend Conv nets (CNNs) [66] to arbitrarily connected undirected graphs. GCNs learn representations for every node in the graph that encodes both the local structure of the graph surrounding the node of interest, as well as the features of the node itself. At a graph convolutional layer, features are aggregated from neighboring nodes and the node itself to produce new output features. By stacking multiple layers, we are able to gather information from nodes further away. GCNs have been applied successfully for graph node classification [36], graph link prediction [67], and zero-shot prediction [68]. Knowledge graphs naturally lend themselves to applications of GCNs owing to the underlying structured interactions between nodes connected by relationships of various types. In this work, given an image and a question about the image, we first identify useful sub-graphs of a large knowledge graph such as DBpedia [32] and then use GCNs to produce representations encoding node and neighborhood features that can be used for answering the question.

Specifically, we propose a model that retrieves the most relevant facts to a question-answer pair based on GloVe features. The sub-graph of facts is passed through a graph convolution network which predicts an answer from these facts. Our approach has the following advantages: 1) Unlike FVQA and AHAB, we avoid the step of query construction and do not use the ground truth visual concept or answer type information which makes it possible to incorporate any fact space into our model. 2) We use GloVe embeddings for retrieving and representing facts which works well with synonyms and homographs. 3) In contrast to STTF, which uses a deep network to arrive at the right fact, we use a GCN which operates on a subgraph of relevant facts while retaining the graphical structure of the knowledge base which allows for reasoning using message passing. 4) Unlike previous works, we have reduced the reliance on the knowledge of the ground truth fact at training time.

CHAPTER 3: LEARNING KNOWLEDGE BASE RETRIEVAL

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

3.1 APPROACH

Our developed approach is outlined in Fig. 3.1. The task at hand is to predict an answer y for a question Q given an image x by using an external knowledge base KB, which consists of a set of facts f_i, *i.e.*, $\text{KB} = \left\{f_1, \ldots, f_{|\text{KB}|}\right\}$. Each fact f_i in the knowledge base is represented as a Resource Description Framework (RDF) triplet of the form $f_i = (a_i, r_i, b_i)$, where a_i is a visual concept in the image, b_i is an attribute or phrase associated with the visual entity a_i, and $r_i \in \mathcal{R}$ is a relation between the two entities. The dataset contains $|\mathcal{R}| = 13$ relations $r \in \mathcal{R} = \{$*Category*, *Comparative*, *HasA*, *IsA*, *HasProperty*, *CapableOf*, *Desires*, *RelatedTo*, *AtLocation*, *PartOf*, *ReceivesAction*, *UsedFor*, *CreatedBy*$\}$. Example triples of the knowledge base in our dataset are (*Umbrella*, *UsedFor*, *Shade*), (*Beach*, *HasProperty*, *Sandy*), (*Elephant*, *Comparative-LargerThan*, *Ant*).

To answer a question Q correctly given an image x, we need to retrieve the right supporting fact and choose the correct entity, *i.e.*, either a or b. Importantly, entity a is always derived from the image and entity b is derived from the fact base. Consequently we refer to this choice as the answer source $s \in \{\text{Image}, \text{KnowledgeBase}\}$. Using this formulation, we can extract the answer y from a predicted fact $\hat{f} = (\hat{a}, \hat{r}, \hat{b})$ and a predicted answer source $\hat{s}$ using

$$y = \begin{cases} \hat{a}, & \text{from } \hat{f} \text{ if } \hat{s} = \text{Image} \\ \hat{b}, & \text{from } \hat{f} \text{ if } \hat{s} = \text{KnowledgeBase} \end{cases}. \tag{3.1}$$

It remains to answer, how to predict a fact $\hat{f}$ and how to infer the answer source $\hat{s}$. The latter is a binary prediction task and we describe our approach below. For the former, we note that the knowledge base contains a large number of facts. We therefore consider it infeasible to search through all the facts $f_i\ \forall i \in \{1, \ldots, |\text{KB}|\}$ using an expensive evaluation based on a deep net. We therefore split this task into two parts: (1) Given a question, we train a network to predict the relation $\hat{r}$, that the question focuses on. (2) Using the predicted relation, $\hat{r}$, we reduce the fact space to those containing only the predicted relation.

Subsequently, to answer the question Q given image x, we only assess the suitability of the facts which contain the predicted relation $\hat{r}$. To assess the suitability, we design a score

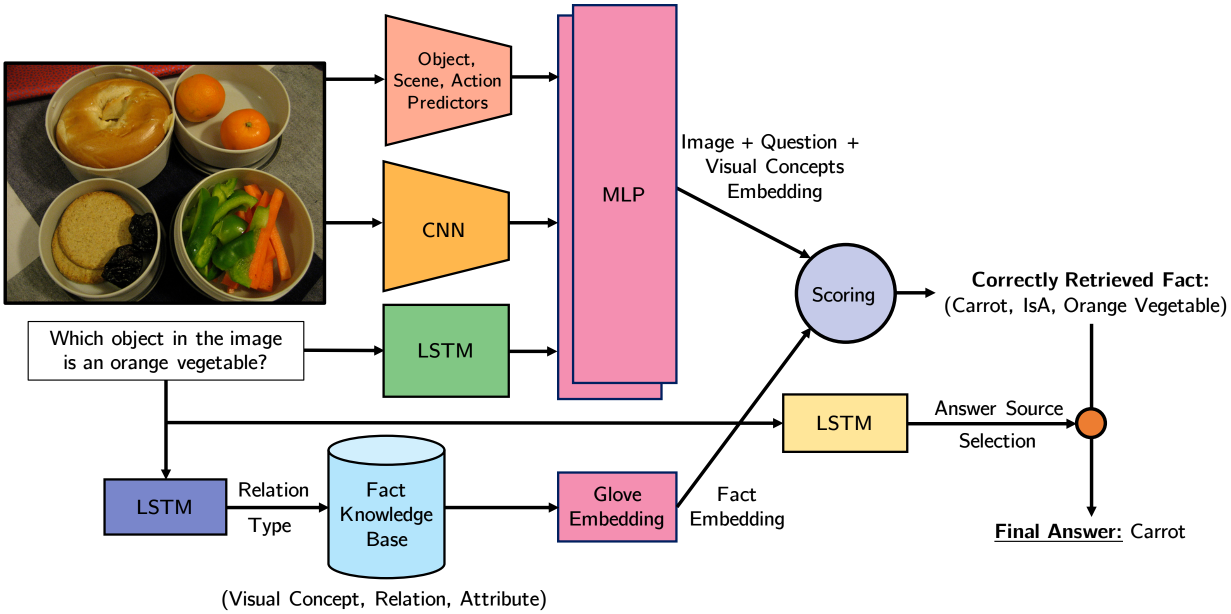

Object, Scene, Action Predictors
CNN
LSTM
MLP
Which object in the image is an orange vegetable?
Image + Question + Visual Concepts Embedding
Scoring
Correctly Retrieved Fact: (Carrot, IsA, Orange Vegetable)
LSTM
Answer Source Selection
LSTM
Relation Type
Fact Knowledge Base
Glove Embedding
Fact Embedding
Final Answer: Carrot
(Visual Concept, Relation, Attribute)

Figure 3.1: Overview of the proposed approach. Given an image and a question about the image, we obtain an Image + Question Embedding through the use of a CNN on the image, an LSTM on the question, and a Multi Layer Perceptron (MLP) for combining the two modalities. In order to filter relevant facts from the Knowledge Base (KB), we use another LSTM to predict the fact relation type from the question. The retrieved structured facts are encoded using GloVe embeddings. The retrieved facts are ranked through a dot product between the embedding vectors and the top-ranked fact is returned to answer the question.

function $S(g^{\mathrm{F}}(f_i), g^{\mathrm{NN}}(x,Q))$ which measures the compatibility of a fact representation $g^{\mathrm{F}}(f_i)$ and an image-question representation $g^{\mathrm{NN}}(x,Q)$. Intuitively, the higher the score, the more suitable the fact f_i for answering question Q given image x.

Formally, we hence obtain the predicted fact $\hat{f}$ via

$$\hat{f} = \arg \max_{i \in \{j : \mathrm{rel}(f_j) = \hat{r}\}} S(g^{\mathrm{F}}(f_i), g^{\mathrm{NN}}(x,Q)), \tag{3.2}$$

where we search for the fact $\hat{f}$ maximizing the score S among all facts f_i which contain relation $\hat{r}$, *i.e.*, among all f_i with $i \in \{j : \mathrm{rel}(f_j) = \hat{r}\}$. Hereby we use the operator $\mathrm{rel}(f_i)$ to indicate the relation of the fact triplet f_i. Given the predicted fact using Eq. (3.2) we obtain the answer y from Eq. (3.1) after predicting the answer source $\hat{s}$.

This approach is outlined in Fig. 3.1. Pictorially, we illustrate the construction of an image-question embedding $g^{\mathrm{NN}}(x,Q)$, via LSTM and CNN net representations that are combined via an MLP. We also illustrate the fact embedding $g^{\mathrm{F}}(f_i)$. Both of them are combined using the score function $S(\cdot,\cdot)$, to predict a fact $\hat{f}$ from which we extract the answer as described in Eq. (3.1).

In the following, we first provide details about the score function S, before discussing prediction of the relation $\hat{r}$ and prediction of the answer source $\hat{s}$.

3.1.1 Scoring the facts

Fig. 3.1 illustrates our approach to score the facts in the knowledge base, *i.e.*, to compute $S(g^{\mathrm{F}}(f_i), g^{\mathrm{NN}}(x, Q))$. We obtain the score in three steps: (1) computing of a fact representation $g^{\mathrm{F}}(f_i)$; (2) computing of an image-question representation $g^{\mathrm{NN}}(x, Q)$; (3) combination of the fact and image-question representation to obtain the final score S. We discuss each of those steps in the following.

(1) Computing a fact representation

To obtain the fact representation $g^{\mathrm{F}}(f_i)$, we concatenate two vectors, the averaged GloVe-100 [69] representation of the words of entity a_i and the averaged GloVe-100 representation of the words of entity b_i. Note that this fact representation is non-parametric, *i.e.*, there are no trainable parameters.

(2) Computing an image-question representation

We compute the image-question representation $g^{\mathrm{NN}}(x, Q)$, by combining a visual representation $g_w^V(x)$, obtained from a standard deep net, *e.g.*, ResNet or VGG, with a visual concept representation $g_w^C(x)$, and a sentence representation $g_w^Q(Q)$, of the question Q, obtained using a trainable recurrent net. For notational convenience we concatenate all trainable parameters into one vector w. Making the dependence on the parameters explicit, we obtain the image-question representation via $g_w^{\mathrm{NN}}(x, Q) = g_w^{\mathrm{NN}}(g_w^V(x), g_w^Q(Q), g_w^C(x))$.

More specifically, for the question embedding $g_w^Q(Q)$, we use an LSTM model [70]. For the image embedding $g_w^V(x)$, we extract image features using ResNet-152 [71] pre-trained on the ImageNet dataset [72]. In addition, we also extract a visual concept representation $g_w^C(x)$, which is a multi-hot vector of size 1176 indicating the visual concepts which are grounded in the image. The visual concepts detected in the images are objects, scenes, and actions. For *objects*, we use the detections from two Faster-RCNN [73] models that are trained on the Microsoft COCO 80-object [74] and the ImageNet 200-object [75] datasets. In total, there are 234 distinct object classes, from which we use that subset of labels that coincides with the FVQA dataset. The *scene* information (such as pasture, beach, bedroom) is extracted by the

VGG-16 model [**?**] trained on the MIT Places 365-class dataset [76]. Again, we use a subset of Places to construct the 1176-dimensional multi-hot vector $g_w^C(x)$. For detecting *actions*, we use the CNN model proposed in [77] which is trained on the HICO [78] and MPII [79] datasets. The HICO dataset contains labels for 600 human-object interaction activities while the MPII dataset contains labels for 393 actions. We use a subset of actions, namely those which coincide with the ones in the FVQA dataset.

All the three vectors $g_w^V(x), g_w^Q(Q), g_w^C(x)$ are concatenated and passed to the multi-layer perceptron $g_w^{\mathrm{NN}}(\cdot,\cdot,\cdot)$.

(3) Combination of fact and image-question representation

For each fact representation $g^{\mathrm{F}}(f_i)$, we compute a score

$$S_w(g^{\mathrm{F}}(f_i), g_w^{\mathrm{NN}}(x,Q)) = \cos(g^{\mathrm{F}}(f_i), g_w^{\mathrm{NN}}(x,Q)) = \frac{g^{\mathrm{F}}(f_i) \cdot g_w^{\mathrm{NN}}(x,Q)}{||g^{\mathrm{F}}(f_i)|| \cdot ||g_w^{\mathrm{NN}}(x,Q)||},$$

where $g_w^{\mathrm{NN}}(x,Q)$ is the image question representation. Hence, the score S is the cosine similarity between the two normalized representations and represents the fit of fact f_i to the image-question pair (x,Q).

3.1.2 Predicting the relation

To predict the relation $\hat{r} \in \mathcal{R} = h_{w_1}^r(Q)$, from the obtained question Q, we use an LSTM net. More specifically, we first embed and then encode the words of the question Q, one at a time, and linearly transform the final hidden representation of the LSTM to predict $\hat{r}$, from $|\mathcal{R}|$ possibilities using a standard multinomial classification. For the results presented in this work, we trained the relation prediction parameters w_1 independently of the score function. We leave a joint formulation to future work.

3.1.3 Predicting the answer source

Prediction of the answer source $\hat{s} = h_{w_2}^s(Q)$ from a given question Q is similar to relation prediction. Again, we use an LSTM net to embed and encode the words of the question Q before linearly transforming the final hidden representation to predict $\hat{s} \in \{\mathrm{Image}, \mathrm{KnowledgeBase}\}$. Analogous to relation prediction, we train this LSTM net's parameters w_2 separately and leave a joint formulation to future work.

Algorithm 3.1 Training with hard negative mining

```
Input: $(x, Q, f^*)$, $KB$
Output: parameters $w$
for $t = 0, \ldots, T$ do
    Create dataset $\mathcal{D}^{(t)}$ by sampling negative facts randomly (if $t = 0$) or by retrieving facts predicted wrongly with $w^{(t-1)}$ (if $t > 0$)
    Use $\mathcal{D}^{(t)}$ to obtain $w^{(t)}$ by optimizing the program given in Eq. (3.7)
end for
return $w^{(T)}$
```

3.2 LEARNING

As mentioned before, we train the parameters w (score function), w_1 (relation prediction), and w_2 (answer source prediction) separately. To train w_1, we use a dataset $\mathcal{D}_1 = \{(Q, r)\}$ containing pairs of question and the corresponding relation which was used to obtain the answer. To learn w_2, we use a dataset $\mathcal{D}_2 = \{(Q, s)\}$, containing pairs of question and the corresponding answer source. For both classifiers we use stochastic gradient descent on the classical cross-entropy and binary cross-entropy loss respectively. Note that both the datasets are readily available from [30].

To train the parameters of the score function we adopt a successive approach operating in time steps $t = \{1, \ldots, T\}$. In each time step, we gradually increase the difficulty of the dataset $\mathcal{D}^{(t)}$ by mining hard negatives. More specifically, for every question Q, and image x, $\mathcal{D}^{(0)}$ contains the 'groundtruth' fact f^* as well as 99 randomly sampled 'non-groundtruth' facts. After having trained the score function on this dataset we use it to predict facts for image-question pairs and create a new dataset $\mathcal{D}^{(1)}$ which now contains, along with the groundtruth fact, another 99 non-groundtruth facts that the score function assigned a high score to.

Given a dataset $\mathcal{D}^{(t)}$, we train the parameters w of the representations involved in the score function $S_w(g^{\text{F}}(f_i), g_w^{\text{NN}}(x, Q))$, and its image, question, and concept embeddings by encouraging that the score of the groundtruth fact f^* is larger than the score of any other fact. More formally, we aim for parameters w which ensure the classical margin, *i.e.*, an SVM-like loss for deep nets:

$$S_w(f^*, x, Q) \geq L(f^*, f) + S_w(f, x, Q) \qquad \forall (f, x, Q) \in \mathcal{D}^{(t)}, \tag{3.3}$$

where $L(f^*, f)$ is the task loss (aka margin) comparing the groundtruth fact f^* to other facts f. In our case $L \equiv 1$. Since we may not find parameters w which ensure feasibility

$\forall(f,x,Q) \in \mathcal{D}^{(t)}$, we introduce slack variables $\xi_{(f,x,Q)} \geq 0$ to obtain after reformulation:

$$\xi_{(f,x,Q)} \geq L(f^*,f) + S_w(f,x,Q) - S_w(f^*,x,Q) \qquad \forall(f,x,Q) \in \mathcal{D}^{(t)}. \tag{3.4}$$

Instead of enforcing the constraint $\forall(f,x,Q)$ in the dataset $\mathcal{D}^{(t)}$, it is equivalent to require [80]

$$\xi_{(x,Q)} \geq \max_f \{L(f^*,f) + S_w(f,x,Q)\} - S_w(f^*,x,Q) \qquad \forall(x,Q) \in \mathcal{D}^{(t)}. \tag{3.5}$$

Using this constraint, we find the parameters w by solving

$$\min_{w,\xi_{(x,Q)}\geq 0} \frac{C}{2}\|w\|_2^2 + \sum_{(x,Q)\in\mathcal{D}^{(t)}} \xi_{(x,Q)} \quad \text{s.t. Constraints in Eq. (3.5).} \tag{3.6}$$

For applicability of the standard sub-gradient descent techniques, we reformulate the program given in Eq. (3.6) to read as

$$\min_w \frac{C}{2}\|w\|_2^2 + \sum_{(x,Q)\in\mathcal{D}^{(t)}} \left(\max_f \{L(f^*,f) + S_w(f,x,Q)\} - S_w(f^*,x,Q)\right), \tag{3.7}$$

which can be optimized using standard deep net packages. The proposed approach for learning the parameters w is summarized in Alg. 3.1.

3.3 EXPERIMENTS

In the following, we assess the proposed approach. We first provide details about the proposed dataset before presenting quantitative results for prediction of relations from questions, prediction of answer-source from questions, and prediction of the answer and the supporting fact. We also discuss mining of hard negatives. Finally, we show qualitative results.

3.3.1 Dataset and Knowledge Base

We use the publicly available FVQA dataset [30] and its knowledge base to evaluate our model. This dataset consists of 2,190 images, 5,286 questions, and 4,126 unique facts corresponding to the questions. The knowledge base, consisting of 193,449 facts, were constructed by extracting the top visual concepts for all the images in the dataset and querying for those concepts in the three knowledge bases, WebChild [31], ConceptNet [33],

Method	**Accuracy**	
	@1	**@3**
FVQA [30]	64.94	82.42
Ours	**75.4**	**91.97**

Table 3.1: Accuracy of predicting relations given the question.

Method	**Accuracy**	
	@1	**@3**
Ours	97.3	100.00

Table 3.2: Accuracy of predicting answer source from a given question.

and DBPedia [32]. The dataset consists of 5 train-test folds, and all the scores we report are averaged across all splits.

3.3.2 Predicting Relations from Questions

We use an LSTM architecture as discussed in Sec. 3 to predict the relation $r \in \mathcal{R}$ given a question Q. The standard train-test split of the FVQA dataset is used to evaluate our model. Batch gradient descent with Adam optimizer was used on batches of size 100 and the model was trained over 50 epochs. LSTM embedding and word embeddings are of size 128 each. The learning rate is set to 1e$-$3 and a dropout of 0.7 is applied after the word embeddings as well as the LSTM embedding. Table 3.1 provides a comparison of our model to the FVQA baseline [30] using top-1 and top-3 prediction accuracy. We observe our results to improve the baseline by more than 10% on top-1 accuracy and by more than 9% when using the top-3 accuracy metric.

3.3.3 Predicting Answer Source from Questions

We assess the accuracy of predicting the answer source s given a question Q. To predict the source of the answer, we use an LSTM architecture as discussed in detail in Sec. 3. Note that for predicting the answer source, the size of the LSTM embedding and word embeddings was set to 64 each. Table 3.2 summarizes the accuracy of the prediction results of our model. We observe the prediction accuracy of the proposed approach to be close to perfect.

3.3.4 Predicting the Correct Answer

Our score function based model to retrieve the supporting fact is described in detail in Sec. 3. For the image embedding, we pass the 2048 dimensional feature vector returned by ResNet through a fully-connected layer and reduce it to a 64 dimensional vector. For the question embedding, we use an LSTM with a hidden layer of size 128. The two are then concatenated into a vector of size 192 and passed through a two layer perceptron with 256 and 128 nodes respectively. Note that the baseline doesn't use image features apart from the detected visual concepts.

The multi-hot visual concept embedding is passed through a fully-connected layer to form a 128 dimensional vector. This is then concatenated with the output of the perceptron and passed through another layer with 200 output nodes. We found a late fusion of the visual concepts to results in a better model as the facts explicitly contain these terms.

Fact embeddings are constructed using GloVe-100 vectors each, for entities a and b. If a or b contain multiple words, an average of all the embeddings is computed. We use cosine distance between the MLP and the fact embeddings to score the facts. The highest scoring fact is chosen as the answer. Ties are broken randomly.

Based on the answer source prediction which is computed using the aforementioned LSTM model, we choose either entity a or b of the fact to be the answer. See Eq. (3.1) for the formal description. Accuracy is computed based on exact match between the chosen entity and the groundtruth answer.

To assess the importance of particular features we investigate 5 variants of our model with varying features: two oracle approaches '*gt* Question + Image + Visual Concepts' and '*gt* Question + Visual Concepts' which make use of groundtruth relation type and answer type data. More specifically, '*gt* Question + Image + Visual Concepts' and '*gt* Question + Visual Concepts' use the groundtruth relations and answer sources respectively. We have three approaches using a variety of features as follows: 'Question + Image + Visual Concepts,' 'Question + Visual Concepts,' and 'Question + Image.' We drop either the Image embeddings from ResNet or the Visual Concept embeddings to obtain two other models, 'Question + Visual Concepts' and 'Question + Image.'

Table 3.3 shows the accuracy of our model in predicting an answer and compares our results to other FVQA baselines. We observe the proposed approach to outperform the state-of-the-art ensemble technique by more than 3% and the strongest baseline without ensemble by over 5% on the top-1 accuracy metric. Moreover we note the importance of visual concepts to accurately predict the answer. By including groundtruth information we assess the maximally possible top-1 and top-3 accuracy. We observe the difference to be around 8%, suggesting that there is some room for improvement.

3.3.5 Question to Supporting Fact

To provide a complete assessment of the proposed approach we illustrate in Table 3.4 the top-1 and top-3 accuracy scores in retrieving the supporting facts of our model compared to other FVQA baselines. We observe the proposed approach to improve significantly both the top-1 and top-3 accuracy by more than 20%. We think this is a significant improvement towards efficiently including knowledge bases into visual question answering.

Method	**Accuracy**	
	@1	**@3**
LSTM-Question+Image+Pre-VQA [30]	24.98	40.40
Hie-Question+Image+Pre-VQA [30]	43.14	59.44
FVQA [30]	56.91	64.65
Ensemble [30]	58.76	-
Ours - Question + Image	26.68	30.27
Ours - Question + Image + Visual Concepts	60.30	73.10
Ours - Question + Visual Concepts	**62.20**	**75.60**
Ours - *gt* Question + Image + Visual Concepts	69.12	80.25
Ours - *gt* Question + Visual Concepts	70.34	82.12

Table 3.3: Answer accuracy over the FVQA dataset.

Method	**Accuracy**	
	@1	**@3**
FVQA-top-1 [30]	38.76	42.96
FVQA-top-3 [30]	41.12	45.49
Ours - Question + Image	28.98	32.34
Ours - Question + Image + Visual Concepts	62.30	74.90
Ours - Question + Visual Concepts	**64.50**	**76.20**

Table 3.4: Correct fact prediction precision over the FVQA dataset.

3.3.6 Mining Hard Negatives

We trained our model over three iterations of hard negative mining, *i.e.*, $T = 2$. In iteration 1 ($t = 0$), all the 193,449 facts were used to sample the 99 negative facts during train. At every 10th epoch of training, negative facts which received high scores were saved. In the

Iteration	# Hard Negatives	Precision	
		@1	@3
1	0	20.17	23.46
2	84,563	38.65	45.49
3	6,889	64.5	76.2

Table 3.5: Correct fact prediction precision with hard negative mining.

next iteration, the trained model along with the negative facts is loaded and we ensure that the 99 negative facts are now sampled from the hard negatives. Table 3.5 shows the Top-1 and Top-3 accuracy for predicting the supporting facts over each of the three iterations. We observe significant improvements due to the proposed hard negative mining strategy. While naïve training of the proposed approach yields only 20.17% top-1 accuracy, two iterations improve the performance to 64.5%.

3.3.7 Synonyms and Homographs

Here we show the improvements of our model compared to the baseline with respect to synonyms and homographs. To this end, we run additional tests using Wordnet to determine the number of question-fact pairs which contain synonyms. The test data contains 1105 such pairs out of which our model predicts 91.6% (1012) correctly, whereas the FVQA model predicts 78.0% (862) correctly. In addition, we manually generated 100 synonymous questions by replacing words in the questions with synonyms (*e.g.*, "What in the bowl can you eat?" is rephrased to "What in the bowl is edible?"). Tests on these 100 new samples find that our model predicts 89 of these correctly, whereas the key-word matching FVQA technique [30] gets 61 of these right. With regards to homographs, the test set has 998 questions which contain words that have multiple meanings across facts. Our model predicts correct answers for 79.4% (792), whereas the FVQA model gets 66.3% (662) correct.

Table 3.6 shows the top-10 accuracy in predicting the answers, averaged across all the test splits of the FVQA dataset. Our model shows an improvement of nearly 11% compared to the strongest baseline, Hie-Question+Image+Pre-VQA [30]. Our models also outperforms by approx. 10% the FVQA gt-QQmapping [30] method, which is the FVQA baseline using the ground truth question queries.

Method	**Accuracy @10**
LSTM-Question+Image+Pre-VQA [30]	57.27
FVQA, top-1 [30]	60.58
FVQA, top-3 [30]	65.54
Hie-Question+Image+Pre-VQA [30]	72.20
FVQA, gt-QQmapping [30]	72.55
Ours - Image + Question	58.96
Ours - Question + Visual Concepts	81.30
Ours - Image + Question + Visual Concepts	**83.92**
Ours - gt Question + Visual Concepts	90.45
Ours - gt Image + Question + Visual Concepts	92.93

Table 3.6: Accuracy at Top-10 in predicting the answer on the FVQA test splits. The best model is indicated in bold.

Method	**KB-Source Accuracy**								
	DBpedia			**ConceptNet**			**WebChild**		
	Top-1	**Top-3**	**Top-10**	**Top-1**	**Top-3**	**Top-10**	**Top-1**	**Top-3**	**Top-10**
LSTM-Question+Image+Pre-VQA [30]	15.38	32.64	51.80	25.97	41.02	57.35	34.42	50.27	68.56
Hie-Question+Image+Pre-VQA [30]	38.39	56.07	71.60	43.23	59.47	72.21	52.85	66.49	73.40
FVQA, top-1 [30]	51.25	63.07	63.13	53.50	60.16	61.20	43.54	46.58	47.03
FVQA, top-3 [30]	56.67	69.31	69.36	57.60	64.70	65.77	48.74	53.45	53.90
Ensemble [30]	57.08	-	-	58.98	-	-	59.77	-	-
FVQA, gt-QQmapping [30]	65.96	78.80	79.43	64.62	71.75	73.17	45.55	48.29	48.86
Ours - Image + Question	25.29	41.46	62.98	34.12	54.17	66.83	42.12	56.76	70.43
Ours - Image + Question + Visual Concepts	69.17	81.24	**86.34**	67.32	74.29	**79.46**	53.45	66.08	**75.89**
Ours - Question + Visual Concepts	**70.29**	**82.46**	84.58	**68.12**	**75.67**	77.86	**54.97**	**67.18**	74.25
Human	74.41	-	-	78.32	-	-	81.95	-	-

Table 3.7: Accuracy in predicting the correct answer based on the knowledge base(KB). Best model is shown in bold.

Table 3.7 shows the top-1, top-3, and top-10 accuracy in predicting the answers on different knowledge bases. Facts from DBpedia have the relation "Category." For example, "Lemon belongs to the category of citric fruits." This is easy to identify in a question and our model can predict the right supporting fact 70.29% of the time. ConceptNet consists of the 11 relations excluding "Category" and "Comparative." Identifying these relations from the question is pretty straightforward and we achieve an accuracy of 68.12% in predicting the answer on questions that use ConceptNet. The errors in both ConceptNet and DBpedia is mainly due to some visual concepts going undetected. WebChild consists of facts with the relation "Comparative," such as "Elephants are stronger than humans." These relations are easy to identify in a question due to the presence of comparative words. However, the overall accuracy in predicting an answer from WebChild is still lower compared to the other KBs. This is because the visual concept and attribute are sometimes reversed in the fact. For example, in the failure cases shown in Fig. 4 of the main submission, the middle example shows a case where the terms are reordered and this causes the answer to be predicted incorrectly.

Overall, our model performs better than the baseline (FVQA top-1), by about 10% to 15%. This steep increase in accuracy is mainly because our model works well with questions containing synonyms and homographs, as we learn the right facts using embeddings and not keyword matching. We provide the synonymous questions in the Sec. 5. We illustrate qualitative results below.

Method	**Answer Source Accuracy**					
	Image			**Knowledge Base**		
	Top-1	**Top-3**	**Top-10**	**Top-1**	**Top-3**	**Top-10**
LSTM-Question+Image+Pre-VQA [30]	28.97	46.62	65.83	6.13	10.94	16.73
Hie-Question+Image+Pre-VQA [30]	49.93	68.21	82.08	11.61	18.69	26.25
FVQA, top-1 [30]	61.11	68.31	68.34	12.12	19.13	23.91
FVQA, top-3 [30]	66.32	74.11	74.15	12.39	19.87	24.81
Ensemble [30]	68.15	-	-	15.18	-	-
FVQA, gt-QQmapping [30]	73.69	81.04	81.04	15.95	25.17	32.39
Ours - Image + Question	31.38	47.52	67.80	10.42	14.67	20.20
Ours - Image + Question + Visual Concepts	75.68	83.14	**86.03**	28.75	38.99	**48.40**
Ours - Question + Visual Concepts	**76.32**	**84.48**	85.94	**29.40**	**39.62**	47.81
Human	82.97	-	-	54.47	-	-

Table 3.8: Accuracy in predicting the correct answer based on the answer source. Best model is shown in bold.

Table 3.8 shows the top-1, top-3, and top-10 accuracy for different methods according to the two different answer sources: Image and Knowledge Base (KB). The answer source is "Image" when the answer is a visual concept, and "KB" when the answer is from the fact. Generating the answer from the large fact space is challenging and hence the accuracy with "KB" is less compared to the accuracy with "Image." However, as our model isn't based on keyword search, we are able to outperform the state-of-the-art model by approx. 15% on "KB" answers and approx. 10% on "Image" answers.

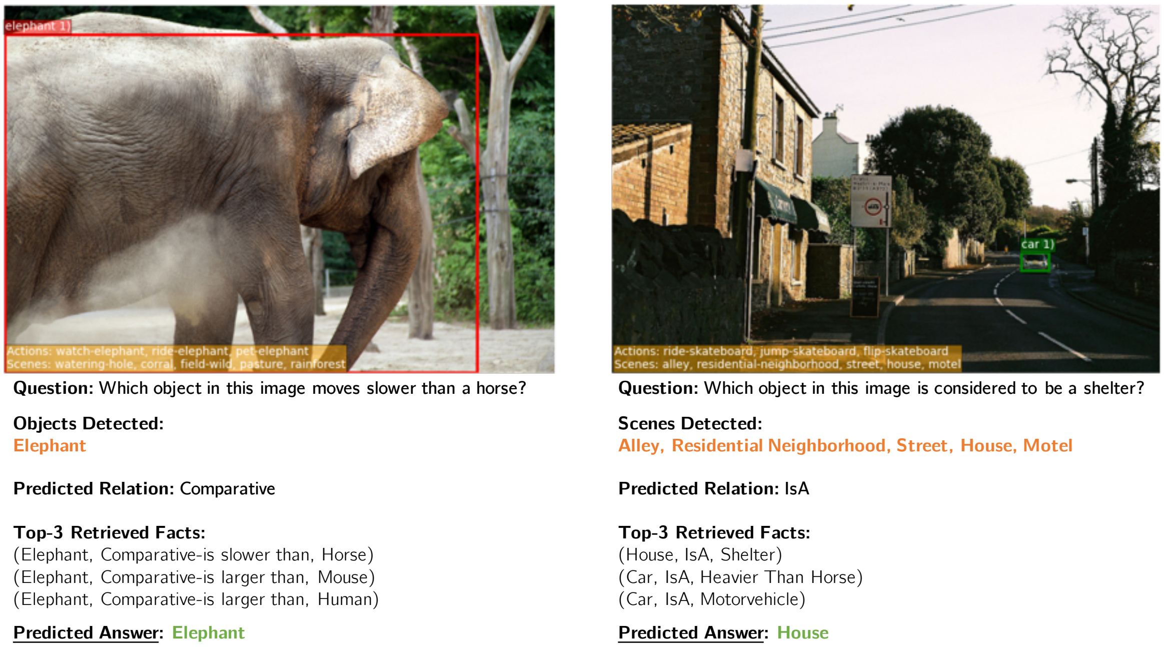
We also report the Wu-Palmer Similarity (WUPS) scores [81] in Tables 4.5 and 4.6. WUPS computes the similarity between two words based on their common subsequence in the taxonomy tree. The predicted answer is considered to be correct if the similarity is greater than a certain threshold. Here we report the WUPS at thresholds 0.9 and 0.0. Our model performs better than the state-of-the-art models at both thresholds.

Method	**WUPS@0.0**		
	@1	**@3**	**@10**
LSTM-Question+Image+Pre-VQA [30]	63.42	76.63	84.94
FVQA, top-1 [30]	64.96	69.57	70.64
Hie-Question+Image+Pre-VQA [30]	71.51	82.71	89.01
FVQA, top-3 [30]	72.34	77.52	78.69
FVQA, gt-QQmapping [30]	73.98	78.67	79.98
Ours - Image + Question	68.55	82.03	89.17
Ours - Image + Question + Visual Concepts	76.10	84.45	**91.68**
Ours - Question + Visual Concepts	**77.25**	**85.27**	90.97
Ours - gt Image + Question + Visual Concepts	81.34	90.59	96.98
Ours - gt Question + Visual Concepts	82.72	91.66	96.15
Human	87.30	-	-

Table 3.9: WUPS@0.0 over the FVQA dataset.

Method	WUPS@0.9		
	@1	@3	@10
LSTM-Question+Image+Pre-VQA [30]	31.96	48.55	64.73
Hie-Question+Image+Pre-VQA [30]	48.93	64.75	76.73
FVQA, top-1 [30]	54.79	61.41	62.22
FVQA, top-3 [30]	59.67	66.89	67.77
FVQA, gt-QQmapping [30]	65.51	72.37	73.55
Ours - Image + Question	38.29	55.14	72.12
Ours - Image + Question + Visual Concepts	68.19	73.23	**80.40**
Ours - Question + Visual Concepts	**69.93**	**75.56**	79.44
Ours - gt Image + Question + Visual Concepts	74.80	83.98	87.12
Ours - gt Question + Visual Concepts	76.34	85.76	89.05
Human	82.47	-	-

Table 3.10: WUPS@0.9 over the FVQA dataset.


elephant 1)
Actions: watch-elephant, ride-elephant, pet-elephant
Scenes: watering-hole, corral, field-wild, pasture, rainforest
car 1)
Actions: ride-skateboard, jump-skateboard, flip-skateboard
Scenes: alley, residential-neighborhood, street, house, motel

Question: Which object in this image moves slower than a horse?

Objects Detected:
Elephant

Predicted Relation: Comparative

Top-3 Retrieved Facts:
(Elephant, Comparative-is slower than, Horse)
(Elephant, Comparative-is larger than, Mouse)
(Elephant, Comparative-is larger than, Human)

Predicted Answer: Elephant

Question: Which object in this image is considered to be a shelter?

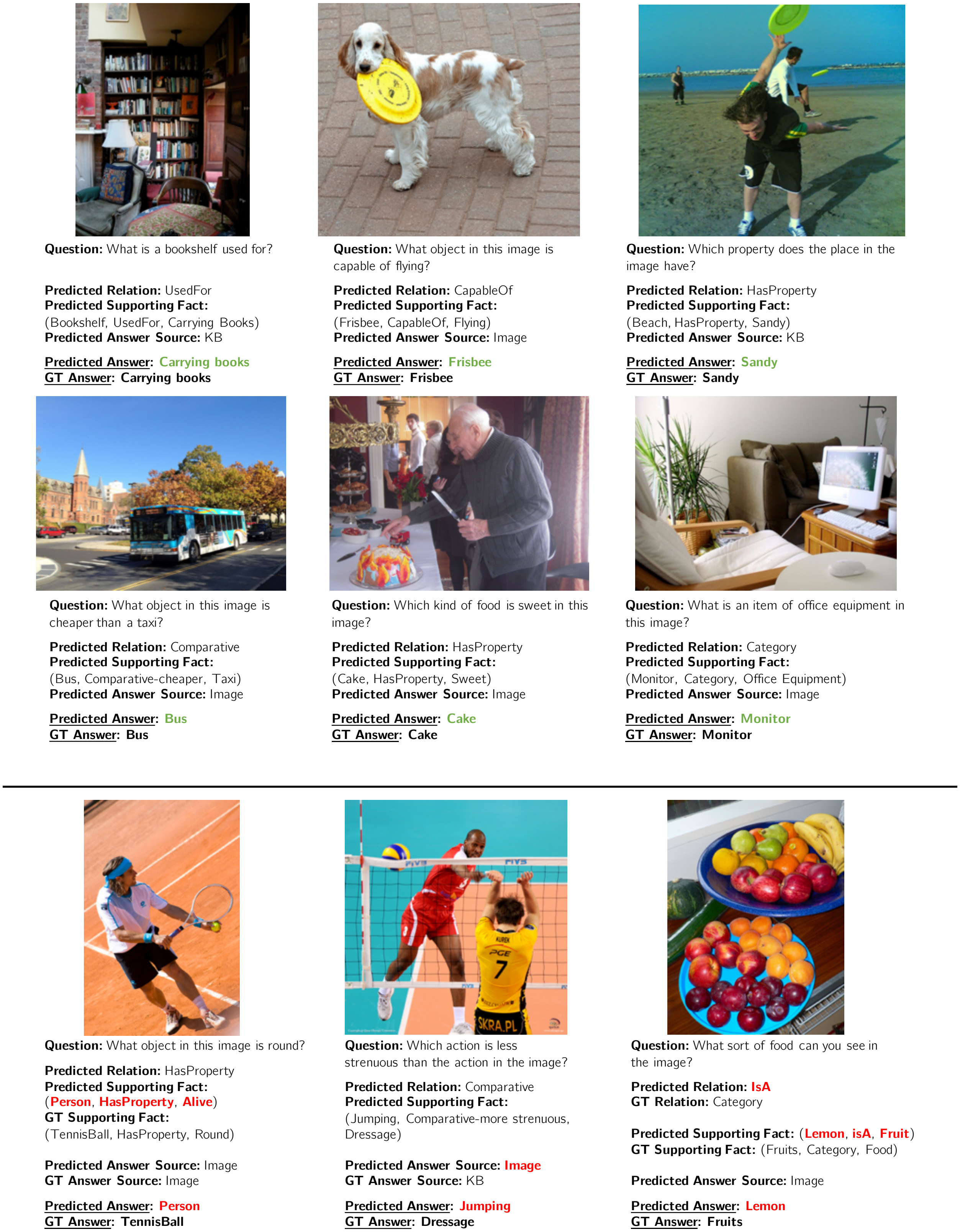
Scenes Detected:
Alley, Residential Neighborhood, Street, House, Motel

Predicted Relation: IsA

Top-3 Retrieved Facts:
(House, IsA, Shelter)
(Car, IsA, Heavier Than Horse)
(Car, IsA, Motorvehicle)

Predicted Answer: House

Figure 3.2: Examples of Visual Concepts (VCs) detected by our framework. Here, we show examples of detected objects, scenes, and actions predicted by the various networks used in our pipeline. There is a clear alignment between useful facts, and the predicted VCs. As a result, including VCs in our scoring method helps improve performance.



Question: What is a bookshelf used for?

Predicted Relation: UsedFor
Predicted Supporting Fact:
(Bookshelf, UsedFor, Carrying Books)
Predicted Answer Source: KB

Predicted Answer: Carrying books
GT Answer: Carrying books

Question: What object in this image is capable of flying?

Predicted Relation: CapableOf
Predicted Supporting Fact:
(Frisbee, CapableOf, Flying)
Predicted Answer Source: Image

Predicted Answer: Frisbee
GT Answer: Frisbee

Question: Which property does the place in the image have?

Predicted Relation: HasProperty
Predicted Supporting Fact:
(Beach, HasProperty, Sandy)
Predicted Answer Source: KB

Predicted Answer: Sandy
GT Answer: Sandy

Question: What object in this image is cheaper than a taxi?

Predicted Relation: Comparative
Predicted Supporting Fact:
(Bus, Comparative-cheaper, Taxi)
Predicted Answer Source: Image

Predicted Answer: Bus
GT Answer: Bus

Question: Which kind of food is sweet in this image?

Predicted Relation: HasProperty
Predicted Supporting Fact:
(Cake, HasProperty, Sweet)
Predicted Answer Source: Image

Predicted Answer: Cake
GT Answer: Cake

Question: What is an item of office equipment in this image?

Predicted Relation: Category
Predicted Supporting Fact:
(Monitor, Category, Office Equipment)
Predicted Answer Source: Image

Predicted Answer: Monitor
GT Answer: Monitor

Question: What object in this image is round?

Predicted Relation: HasProperty
Predicted Supporting Fact:
(**Person**, **HasProperty**, **Alive**)
GT Supporting Fact:
(TennisBall, HasProperty, Round)

Predicted Answer Source: Image
GT Answer Source: Image

Predicted Answer: Person
GT Answer: TennisBall

Question: Which action is less strenuous than the action in the image?

Predicted Relation: Comparative
Predicted Supporting Fact:
(Jumping, Comparative-more strenuous, Dressage)

Predicted Answer Source: Image
GT Answer Source: KB

Predicted Answer: Jumping
GT Answer: Dressage

Question: What sort of food can you see in the image?

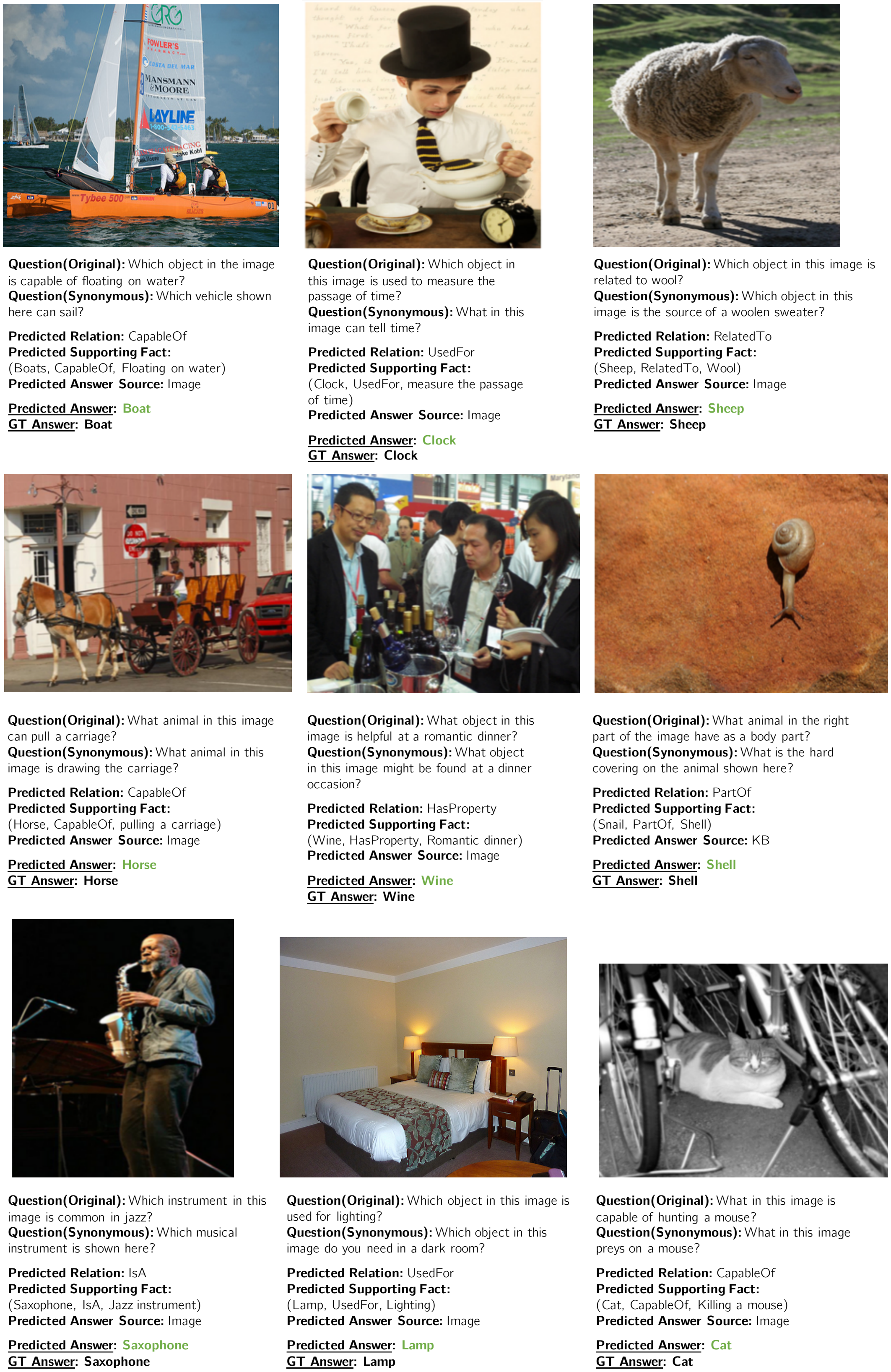
Predicted Relation: IsA
GT Relation: Category

Predicted Supporting Fact: (**Lemon**, **isA**, **Fruit**)
GT Supporting Fact: (Fruits, Category, Food)

Predicted Answer Source: Image

Predicted Answer: Lemon
GT Answer: Fruits

Figure 3.3: Success and failure cases of our method. In the top two rows, our method correctly predicts the relation, the supporting fact, and the answer source to produce the correct answer for the given question. The bottom row of examples shows the failure modes of our method.



Question(Original): Which object in the image is capable of floating on water?
Question(Synonymous): Which vehicle shown here can sail?

Predicted Relation: CapableOf
Predicted Supporting Fact:
(Boats, CapableOf, Floating on water)
Predicted Answer Source: Image

Predicted Answer: **Boat**
GT Answer: **Boat**

Question(Original): Which object in this image is used to measure the passage of time?
Question(Synonymous): What in this image can tell time?

Predicted Relation: UsedFor
Predicted Supporting Fact:
(Clock, UsedFor, measure the passage of time)
Predicted Answer Source: Image

Predicted Answer: **Clock**
GT Answer: **Clock**

Question(Original): Which object in this image is related to wool?
Question(Synonymous): Which object in this image is the source of a woolen sweater?

Predicted Relation: RelatedTo
Predicted Supporting Fact:
(Sheep, RelatedTo, Wool)
Predicted Answer Source: Image

Predicted Answer: **Sheep**
GT Answer: **Sheep**

Question(Original): What animal in this image can pull a carriage?
Question(Synonymous): What animal in this image is drawing the carriage?

Predicted Relation: CapableOf
Predicted Supporting Fact:
(Horse, CapableOf, pulling a carriage)
Predicted Answer Source: Image

Predicted Answer: **Horse**
GT Answer: **Horse**

Question(Original): What object in this image is helpful at a romantic dinner?
Question(Synonymous): What object in this image might be found at a dinner occasion?

Predicted Relation: HasProperty
Predicted Supporting Fact:
(Wine, HasProperty, Romantic dinner)
Predicted Answer Source: Image

Predicted Answer: **Wine**
GT Answer: **Wine**

Question(Original): What animal in the right part of the image have as a body part?
Question(Synonymous): What is the hard covering on the animal shown here?

Predicted Relation: PartOf
Predicted Supporting Fact:
(Snail, PartOf, Shell)
Predicted Answer Source: KB

Predicted Answer: **Shell**
GT Answer: **Shell**

Question(Original): Which instrument in this image is common in jazz?
Question(Synonymous): Which musical instrument is shown here?

Predicted Relation: IsA
Predicted Supporting Fact:
(Saxophone, IsA, Jazz instrument)
Predicted Answer Source: Image

Predicted Answer: **Saxophone**
GT Answer: **Saxophone**

Question(Original): Which object in this image is used for lighting?
Question(Synonymous): Which object in this image do you need in a dark room?

Predicted Relation: UsedFor
Predicted Supporting Fact:
(Lamp, UsedFor, Lighting)
Predicted Answer Source: Image

Predicted Answer: **Lamp**
GT Answer: **Lamp**

Question(Original): What in this image is capable of hunting a mouse?
Question(Synonymous): What in this image preys on a mouse?

Predicted Relation: CapableOf
Predicted Supporting Fact:
(Cat, CapableOf, Killing a mouse)
Predicted Answer Source: Image

Predicted Answer: **Cat**
GT Answer: **Cat**

Figure 3.4: Synonymous and original questions which were answered correctly.

Question: What kind of sport do people usually practice in this place?

Predicted Relation: AtLocation
Predicted Supporting Fact:
(Skiing, AtLocation, Ski-slope)
Predicted Answer Source: KB

Predicted Answer: Skiing
GT Answer: Skiing

Question: Which object in the image is used to make a cake?

Predicted Relation: UsedFor
Predicted Supporting Fact:
(Oven, UsedFor, Baking)
Predicted Answer Source: Image

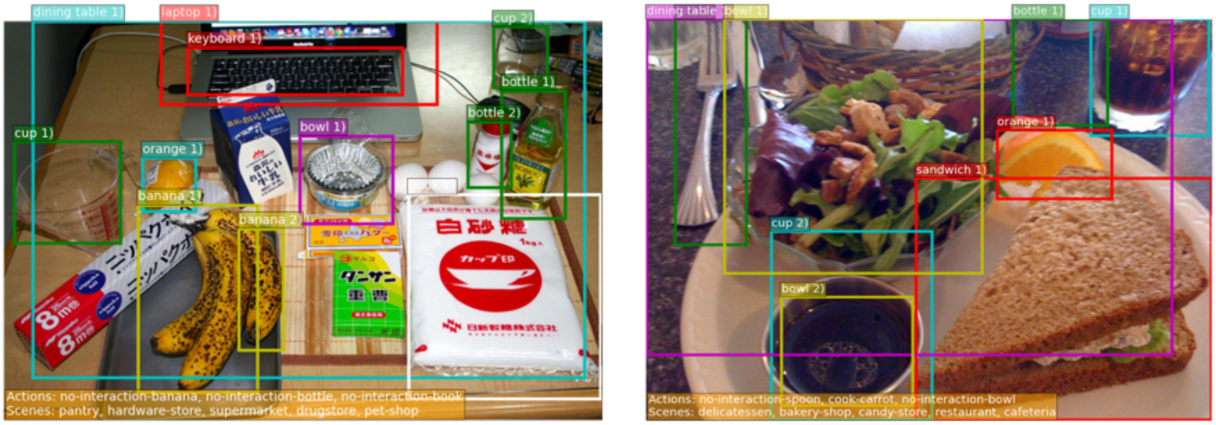
Predicted Answer: Oven
GT Answer: Oven

Question(Original): Which object in this image is related to sailing?

Predicted Relation: RelatedTo
Predicted Supporting Fact:
(Boat, RelatedTo, Sail)
Predicted Answer Source: Image

Predicted Answer: Boat
GT Answer: Boat

Figure 3.5: Questions in test which were predicted incorrectly by the baseline but correctly by our model.


dining table 1)
laptop 1)
keyboard 1)
cup 2)
bottle 1)
bottle 2)
bowl 1)
cup 1)
orange 1)
banana 1)
banana 2)
Actions: no-interaction-banana, no-interaction-bottle, no-interaction-book
Scenes: pantry, hardware-store, supermarket, drugstore, pet-shop
dining table
bowl 1)
bottle 1)
cup 1)
orange 1)
sandwich 1)
cup 2)
bowl 2)
Actions: no-interaction-spoon, cook-carrot, no-interaction-bowl
Scenes: delicatessen, bakery-shop, candy-store, restaurant, cafeteria

Question: Which object are you likely to find in a monkey's hand?

Objects Detected:
Banana, Bowl, Cup, Bottle, Laptop, Keyboard, Dining Table. Book, Orange

Predicted Relation: AtLocation

Top Retrieved Facts:
(Bananas, AtLocation, monkey's hand)
(Banana's, AtLocation, Grocery store)
(Cup, AtLocation, Kitchen)

Predicted Answer: Monkey's Hand

Question: Which object here is rich in Vitamin C?

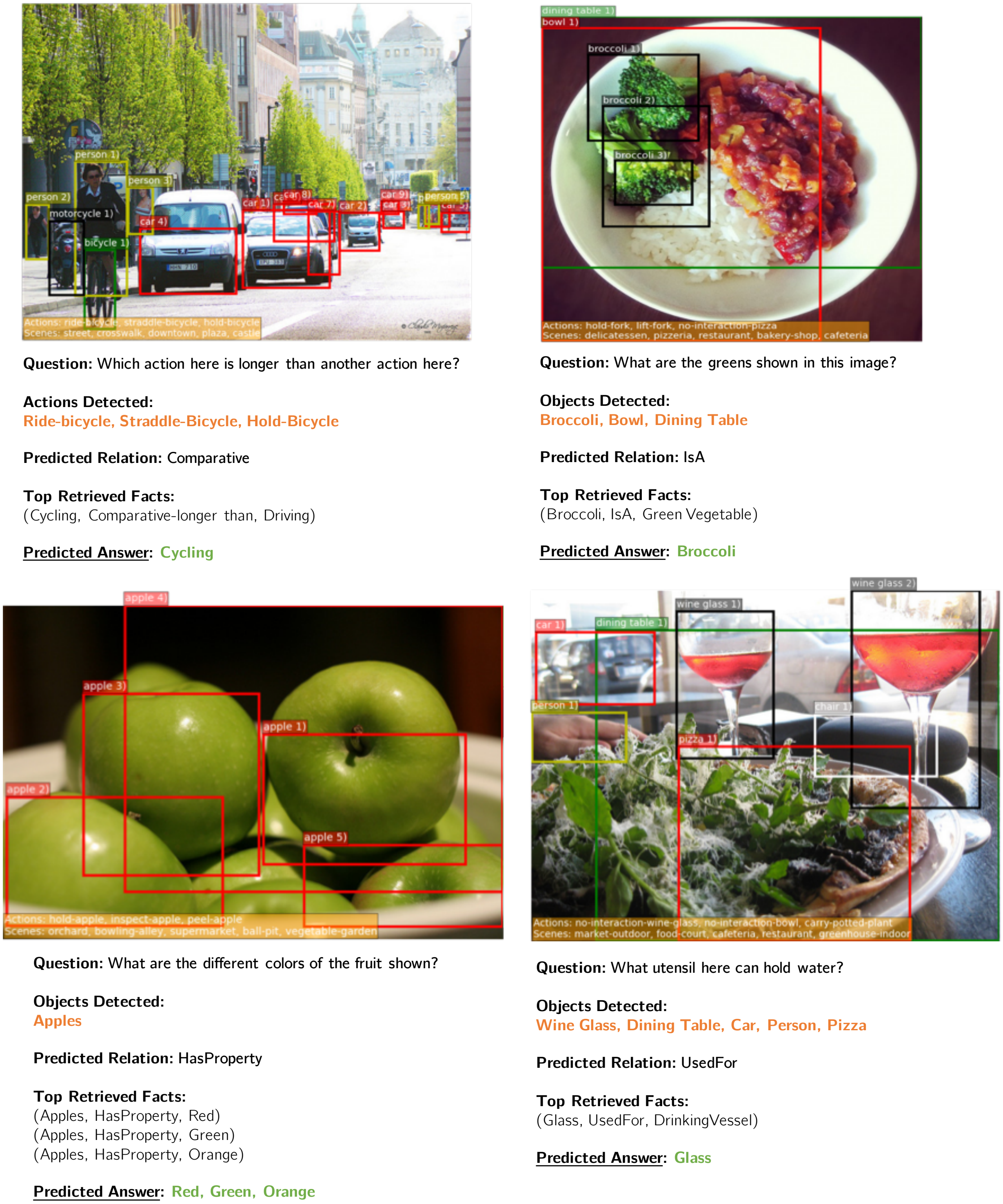
Objects Detected:
Dining Table, Bowl, Bottle, Cup, Sandwich, Fork, Orange

Predicted Relation: HasA

Top Retrieved Facts:
(Orange, HasA, High Vitamin C content)
(Orange juice, HasA, High Vitamin C content)
(Lemon, HasA, Vitamin C)

Predicted Answer: Orange

Figure 3.6: Correct predictions with detected visual concepts and top retrieved facts.


person 1)
person 2)
person 3)
motorcycle 1)
bicycle 1)
car 1)
car 2)
car 3)
car 4)
car 7)
car 8)
car 9)
person 5)
Actions: ride-bicycle, straddle-bicycle, hold-bicycle
Scenes: street, crosswalk, downtown, plaza, castle
dining table 1)
bowl 1)
broccoli 1)
broccoli 2)
broccoli 3)
Actions: hold-fork, lift-fork, no-interaction-pizza
Scenes: delicatessen, pizzeria, restaurant, bakery-shop, cafeteria
apple 1)
apple 2)
apple 3)
apple 4)
apple 5)
Actions: hold-apple, inspect-apple, peel-apple
Scenes: orchard, bowling-alley, supermarket, ball-pit, vegetable-garden
wine glass 1)
wine glass 2)
car 1)
dining table 1)
person 1)
chair 1)
pizza 1)
Actions: no-interaction-wine-glass, no-interaction-bowl, carry-potted-plant
Scenes: market-outdoor, food-court, cafeteria, restaurant, greenhouse-indoor

Question: Which action here is longer than another action here?

Actions Detected:
Ride-bicycle, Straddle-Bicycle, Hold-Bicycle

Predicted Relation: Comparative

Top Retrieved Facts:
(Cycling, Comparative-longer than, Driving)

Predicted Answer: Cycling

Question: What are the greens shown in this image?

Objects Detected:
Broccoli, Bowl, Dining Table

Predicted Relation: IsA

Top Retrieved Facts:
(Broccoli, IsA, Green Vegetable)

Predicted Answer: Broccoli

Question: What are the different colors of the fruit shown?

Objects Detected:
Apples

Predicted Relation: HasProperty

Top Retrieved Facts:
(Apples, HasProperty, Red)
(Apples, HasProperty, Green)
(Apples, HasProperty, Orange)

Predicted Answer: Red, Green, Orange

Question: What utensil here can hold water?

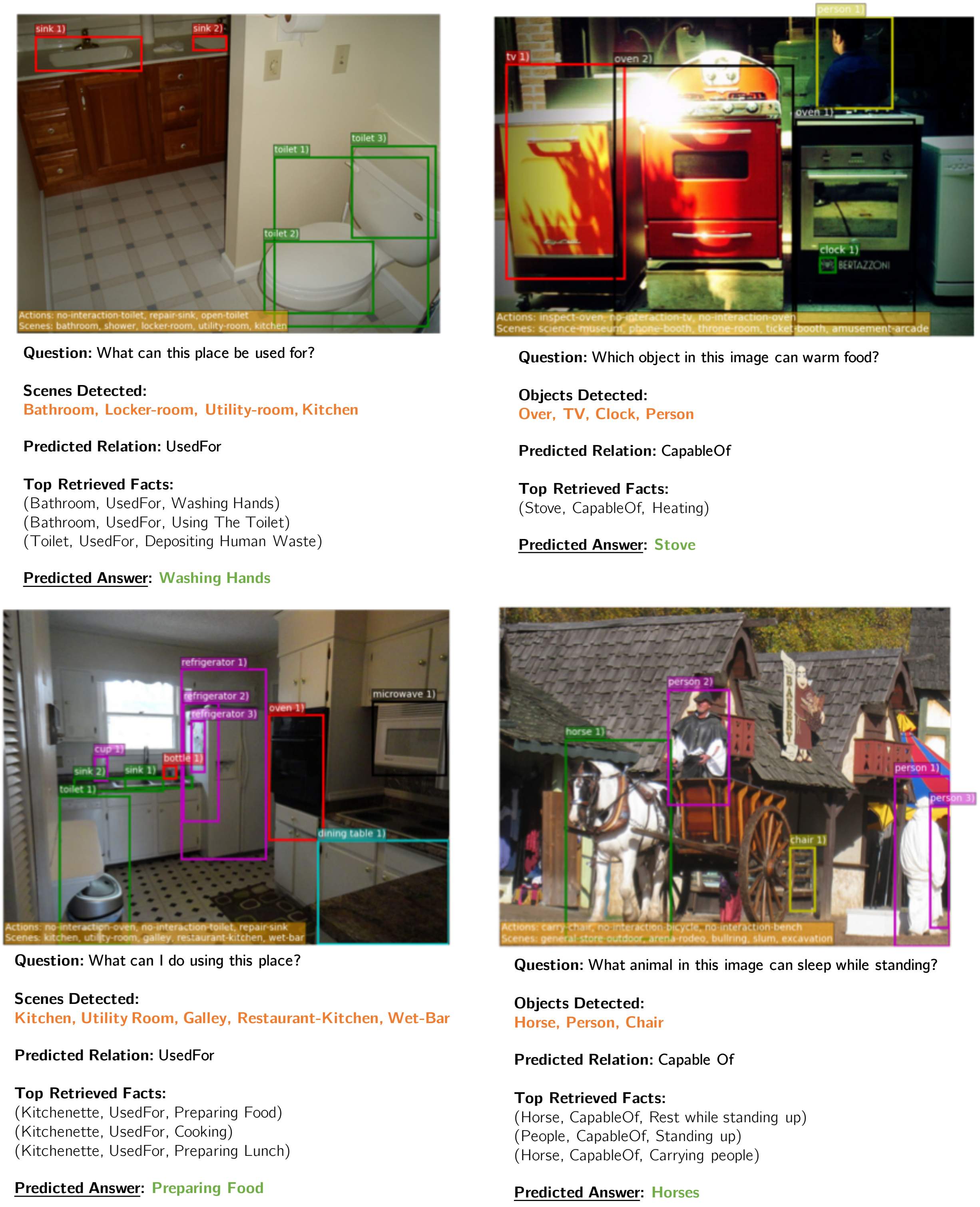
Objects Detected:
Wine Glass, Dining Table, Car, Person, Pizza

Predicted Relation: UsedFor

Top Retrieved Facts:
(Glass, UsedFor, DrinkingVessel)

Predicted Answer: Glass

Figure 3.7: Correct predictions with detected visual concepts and top retrieved facts.


sink 1)
sink 2)
toilet 1)
toilet 2)
toilet 3)
Actions: no-interaction-toilet, repair-sink, open-toilet
Scenes: bathroom, shower, locker-room, utility-room, kitchen
person 1)
tv 1)
oven 2)
oven 1)
clock 1)
Actions: inspect-oven, no-interaction-tv, no-interaction-oven
Scenes: science-museum, phone-booth, throne-room, ticket-booth, amusement-arcade
refrigerator 1)
refrigerator 2)
refrigerator 3)
oven 1)
microwave 1)
cup 1)
bottle 1)
sink 2)
sink 1)
toilet 1)
dining table 1)
Actions: no-interaction-oven, no-interaction-toilet, repair-sink
Scenes: kitchen, utility-room, galley, restaurant-kitchen, wet-bar
person 2)
horse 1)
person 1)
person 3)
chair 1)
Actions: carry-chair, no-interaction-bicycle, no-interaction-bench
Scenes: general-store-outdoor, arena-rodeo, bullring, slum, excavation

Question: What can this place be used for?

Scenes Detected:
Bathroom, Locker-room, Utility-room, Kitchen

Predicted Relation: UsedFor

Top Retrieved Facts:
(Bathroom, UsedFor, Washing Hands)
(Bathroom, UsedFor, Using The Toilet)
(Toilet, UsedFor, Depositing Human Waste)

Predicted Answer: **Washing Hands**

Question: Which object in this image can warm food?

Objects Detected:
Over, TV, Clock, Person

Predicted Relation: CapableOf

Top Retrieved Facts:
(Stove, CapableOf, Heating)

Predicted Answer: **Stove**

Question: What can I do using this place?

Scenes Detected:
Kitchen, Utility Room, Galley, Restaurant-Kitchen, Wet-Bar

Predicted Relation: UsedFor

Top Retrieved Facts:
(Kitchenette, UsedFor, Preparing Food)
(Kitchenette, UsedFor, Cooking)
(Kitchenette, UsedFor, Preparing Lunch)

Predicted Answer: **Preparing Food**

Question: What animal in this image can sleep while standing?

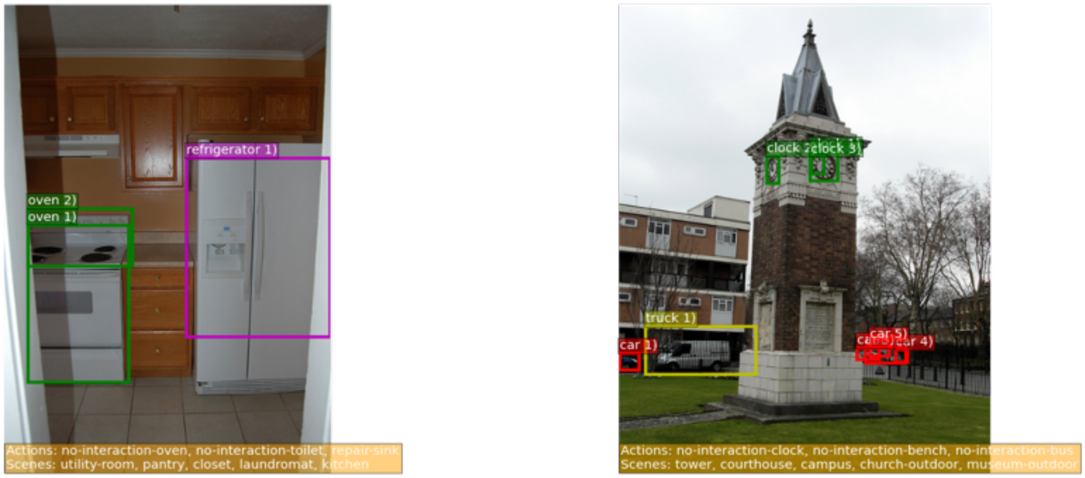
Objects Detected:
Horse, Person, Chair

Predicted Relation: Capable Of

Top Retrieved Facts:
(Horse, CapableOf, Rest while standing up)
(People, CapableOf, Standing up)
(Horse, CapableOf, Carrying people)

Predicted Answer: **Horses**

Figure 3.8: Correct predictions with detected visual concepts and top retrieved facts.


refrigerator 1)
oven 2)
oven 1)
Actions: no-interaction-oven, no-interaction-toilet, repair-sink
Scenes: utility-room, pantry, closet, laundromat, kitchen
clock 2)
clock 3)
truck 1)
car 1)
car 5)
car 4)
Actions: no-interaction-clock, no-interaction-bench, no-interaction-bus
Scenes: tower, courthouse, campus, church-outdoor, museum-outdoor

Question: Which object in the image is utilized to cool food?

Objects Detected:
Oven, Refrigerator

Predicted Relation: UsedFor

Top Retrieved Facts:
(Refrigerator, UsedFor, Chilling the food)
(Refrigerator, UsedFor, Cold food storage)
(Refrigerator, UsedFor, Freezing the food)

Predicted Answer: Refrigerator

Question: Which object in this is used to check the time?

Objects Detected:
Clock, Car, Truck

Predicted Relation: UsedFor

Top Retrieved Facts:
(Clock, UsedFor, Seeing the hour)

Predicted Answer: Clock

Figure 3.9: Correct predictions with detected visual concepts and top retrieved facts.

3.3.8 Qualitative Results.

Fig. 3.2 shows the Visual Concepts (VCs) detected for a few samples along with the top 3 facts retrieved by our model. Providing these predicted VCs as input to our fact-scoring MLP helps improve supporting fact retrieval as well as answer accuracy by a large margin of over 30% as seen in Tables 3.3 and 3.4. As can be seen in Fig. 3.2, there is a close alignment between relevant facts and predicted VCs, as VCs provide a high-level overview of the salient content in the images.

In Fig. 3.3, we show success and failure cases of our method. There are 3 steps to producing the correct answer using our method: (1) correctly predicting the relation, (2) retrieving supporting facts containing the predicted relation, and relevant to the image, and (3) choosing the answer from the predicted answer source (Image/Knowledge Base). The top two rows of images show cases where all the 3 steps were correctly executed by our proposed method. Note that our method works for a variety of relations, objects, answer sources, and varying difficulty. It is correctly able to identify the object of interest, even when it is not the most prominent object in the image. For example, in the middle image of the first row, the frisbee is smaller than the dog in the image. However, we were correctly able to retrieve the supporting fact about the frisbee using information from the question, such as '*capable of*' and '*flying*.'

A mistake in any of the 3 steps can cause our method to produce an incorrect answer. The bottom row of images in Fig. 3.3 displays prototypical failure modes. In the leftmost image, we miss cues from the question such as '*round*,' and instead retrieve a fact about the person. In the middle image, our method makes a mistake at the final step and uses information from the wrong answer source. This is a very rare source of errors overall, as we are over 97% accurate in predicting the answer source, as shown in Table 3.2. In the rightmost image, our method makes a mistake at the first step of predicting the relation, making the remaining steps incorrect. Our relation prediction is around 75%, and 92% accurate by the top-1 and top-3 metrics, as shown in Table 3.1, and has some scope for improvement. For qualitative results regarding synonyms and homographs we refer the interested reader to the supplementary material.

Correctly predicted answers

In this section, we show examples which were correctly predicted by our model but incorrectly predicted by the FVQA baseline model [30].

Fig. 3.5 shows three examples where the baseline model failed to predict the correct answer, but ours succeeded. In the first example, we were able to correctly identify the ski-slope which the baseline couldn't identify. In the second example, the baseline fails as they use keyword matching to retrieve the fact and "make a cake" doesn't equate to the synonym "baking." In the third, the baseline method predicted the relation incorrectly.

In Fig. 3.4, we generated questions synonymous to the ones present in the test set, such that the relation, answer source and the answer are no longer obvious. Our model was able to successfully predict the correct answer for these newly generated complex questions whereas the baseline model was challenged.

In the first two examples, we merely replaced phrases such as "floating on water" and "measure the passage of time" with synonymous phrases "sail" and "tell time" respectively. Our model, capable of handling synonyms, produced the right results on both sets of questions whereas the baseline model which is susceptible to keyword changes suffered.

In the fourth example, we change the phrase "pull a carriage" to "draw a carriage" and our model is able to correctly map the homograph "draw" to "pull." In the sixth example, we remove the term Jazz instrument and replace it with musical instrument. Similarly, in the other examples we replace specific keywords and catchphrases with analogous terms and this causes the baseline model to predict incorrectly, while our model produced the correct results.

Predicted visual concepts and retrieved facts

Figures 3.6 to 3.9 show the Visual Concepts detected for each image along with the top retrieved facts. In all these examples, our simpler model assigns the highest score to the ground truth fact and predicts the answer correctly, whereas the FVQA baseline method [30] produced wrong answers. The FVQA method predicted wrongly in all these cases either because certain visual concepts go undetected or the wrong fact is matched in the keyword matching step. For example, in the second example in Fig. 3.6, the answer predicted by the baseline is Lemon as it wrongly predicts Orange as Lemon and assigns a higher similarity score to the third fact based on the matched words. Our method detects Orange as a visual concept and retrieves the right supporting fact.

In example 2 of Fig. 3.7, our model correctly maps “greens” to “green vegetables” and in example 3 it correctly associates “red, green and yellow” with colors of the fruit. In example 4 of Fig. 3.7, our model is able to interpret the term, “utensil” and match it to the fact about a drinking vessel. Similarly, in Fig. 3.8 example 1, the baseline model fails to detect the scene “bathroom” and wrongly predicts it as “kitchen.” In example 2 and 4 of Fig. 3.8, the baseline predicts a wrong fact as it cannot match the keywords “warm” to “heat” or “sleep” to “rest.”

CHAPTER 4: KNOWLEDGE BASE REASONING WITH GRAPH CONVOLUTIONAL NETWORKS

To jointly 'reason' about a set of answers for a given question-image pair, we develop a graph convolution net (GCN) based approach for visual question answering with knowledge bases. In the following we first provide an overview of the proposed approach before delving into details of the individual components.

4.1 APPROACH

Our proposed approach is outlined in Fig. 4.1. Given an image I and a corresponding question Q, the task is to predict an answer A while using an external knowledge base KB which consists of facts, f_i, *i.e.*, KB $= \{f_1, f_2, \ldots, f_{|\text{KB}|}\}$. A fact is represented as a Resource Distribution Framework (RDF) triplet of the form $f = (x, r, y)$, where x is a visual concept grounded in the image, y is an attribute or phrase, and $r \in \mathcal{R}$ is a relation between the two entities, x and y. The relations in the knowledge base are part of a set of 13 possible relations $\mathcal{R} = \{$*Category*, *Comparative*, *HasA*, *IsA*, *HasProperty*, *CapableOf*, *Desires*, *RelatedTo*, *AtLocation*, *PartOf*, *ReceivesAction*, *UsedFor*, *CreatedBy*$\}$. Subsequently we use $x(f)$, $y(f)$, or $\text{rel}(f)$ to extract the visual concept x, the attribute phrase y, or the relation r in fact $f = (x, r, y)$ respectively.

Every question Q is associated with a single fact, f^*, that helps answer the question. More specifically, the answer A is one of the two entities of that fact, *i.e.*, either $A = x^*$ or $A = y^*$, both of which can be extracted from $f^* = (x^*, r^*, y^*)$.

Wang *et al.* [30] formulate the task as prediction of a fact $\hat{f} = (\hat{x}, \hat{r}, \hat{y})$ for a given question-image pair, and subsequently extract either $\hat{x}$ or $\hat{y}$, depending on the result of an answer source classifier. As there are over $190,000$ facts, retrieving the correct supporting fact f^* is challenging and computationally inefficient. Usage of question properties like 'visual concept type' makes the proposed approach hard to extend.

Guided by the observation that the correct supporting fact f^* is within the top-100 of a retrieval model 84.8% of the time, we develop a two step solution: (1) retrieving the most relevant facts for a given question-image pair. To do this, we extract the top-100 facts, *i.e.*, f_{100} based on word similarity between the question and the fact. Further, we obtain the set of relevant facts f_{rel} by reducing f_{100} based on consistency of the fact relation r with a predicted relation $\hat{r}$. (2) predicting the answer as one of the entities in this reduced fact space f_{rel}. To predict the answer we use a GCN to compute representations of nodes in a graph, where the nodes correspond to the unique entities $e \in E = \{x(f) : f \in f_{\text{rel}}\} \cup \{y(f) : f \in f_{\text{rel}}\}$, *i.e.*,

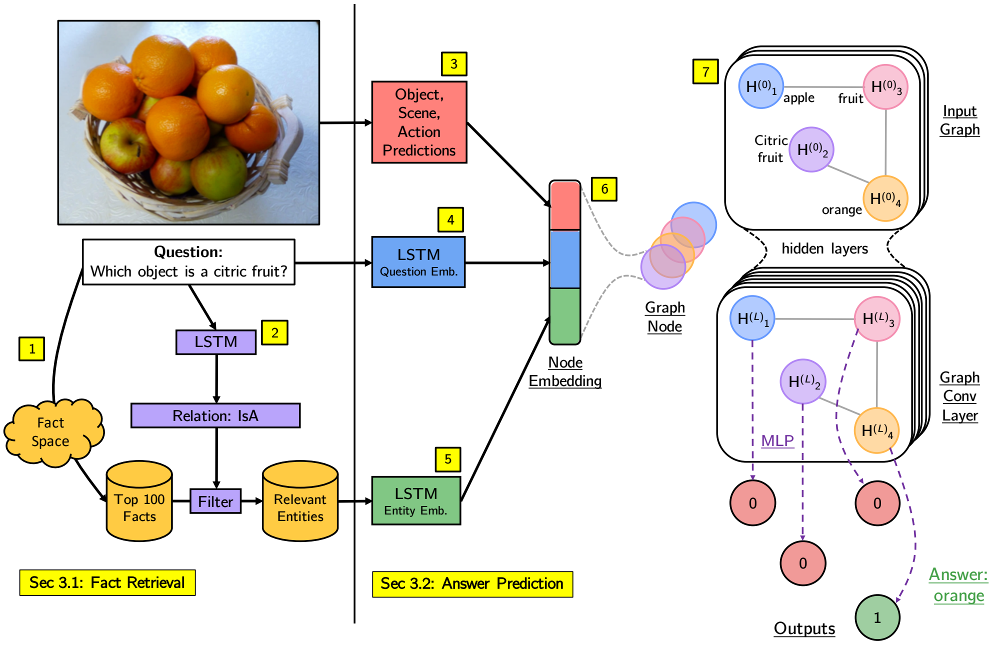


Figure 4.1: Outline of the proposed approach: Given an image and a question, we use a similarity scoring technique (1) to obtain relevant facts from the fact space. An LSTM (2) predicts the relation from the question to further reduce the set of relevant facts and its entities. An entity embedding is obtained by concatenating the visual concepts embedding of the image (3), the LSTM embedding of the question (4), and the LSTM embedding of the entity (5). Each entity forms a single node in the graph and the relations constitute the edges (6). A GCN followed by an MLP performs joint assessment (7) to predict the answer. Our approach is trained end-to-end.

either x or y in the fact space f_{rel}. Two entities in the graph are connected if a fact relates the two. Using a GCN permits to jointly assess the suitability of all entities which makes our proposed approach different from classification based techniques.

For example, consider the image and the question shown in Fig. 4.1. The relation for this question is “IsA” and the fact associated with this question-image pair is (Orange, IsA, Citric). The answer is Orange. In the following we first discuss retrieval of the most relevant facts for a given question-image pair before detailing our GCN approach for extracting the answer from this reduced fact space.

4.1.1 Retrieval of Relevant Facts

To retrieve a set of relevant facts f_{rel} for a given question-image pair, we pursue a score based approach. We first compute the cosine similarity of the GloVe embeddings of the words in the fact with the words in the question and the words of the visual concepts detected in the image. Because some words may differ between question and fact, we obtain a fact score by averaging the Top-K word similarity scores. We rank the facts based on their similarity and retrieve the top-100 facts for each question, which we denote f_{100}. We chose 100 facts as this gives the best downstream accuracy as shown in Tab. 4.1. As indicated in Tab. 4.1, we observe a high recall of the ground truth fact in the retrieved facts while using this technique. This motivates us to avoid a complex model which finds the right fact, as used in [30] and [34], and instead use the retrieved facts to directly predict the answer.

We further reduce this set of 100 facts by assessing their relation attribute. To predict the relation from a given question, we use the approach described in [34]. We retain the facts among the top-100 only if their relation agrees with the predicted relation $\hat{r}$, *i.e.*, $f_{\text{rel}} = \{f \in f_{100} : \text{rel}(f) = \hat{r}\}$.

For every question, unique entities in the facts f_{rel} are grouped into a set of candidate entities, $E = \{x(f) : f \in f_{\text{rel}}\} \cup \{y(f) : f \in f_{\text{rel}}\}$, with $|E| \leq 200$ (2 entities/fact and at most 100 facts).

Currently, we train the relation predictor's parameters independently of the remaining model. In future work we aim for an end-to-end model which includes this step.

4.1.2 Answer Prediction

Given the set of candidate entities E, we want to 'reason' about the answer, *i.e.*, we want to predict an entity $\hat{e} \in E$. To jointly assess the suitability of all candidate entities in E, we develop a Graph-Convolution Net (GCN) based approach which is augmented by a multi-layer perceptron (MLP). The nodes in the employed graph correspond to the available entities $e \in E$ and their node representation is given as an input to the GCN. The GCN combines entity representations in multiple iterative steps. The final transformed entity representations learned by the GCN are then used as input in an MLP which predicts a binary label, *i.e.*, $\{1, 0\}$, for each entity $e \in E$, indicating if e is or isn't the answer.

More formally, the goal of the GCN is to learn how to combine representations for the nodes $e \in E$ of a graph, $\mathcal{G} = (E, \mathcal{E})$. Its output feature representations depend on: (1) learnable weights; (2) an adjacency matrix A_{adj} describing the graph structure $\mathcal{E}$. We consider two entities to be connected if they belong to the same fact; (3) a parametric input representation $g_w(e)$ for every node $e \in E$ of the graph. We subsume the original feature representations of all nodes in an $|E| \times D$-dimensional feature matrix $H^{(0)} \in \mathbb{R}^{|E| \times D}$, where D is the number of features. In our case, each node $e \in E$ is represented by the concatenation of the corresponding image, question and entity representation, *i.e.*, $g_w(e) = (g_w^V(I), g_w^Q(Q), g_w^C(e))$. Combining the three representations ensures that each node/entity depends on the image and the question. The node representation is discussed in detail below.

	@1	**@50**	**@100**	**@150**	**@200**	**@500**
Fact Recall	22.6	76.5	84.8	88.4	91.6	93.1
Downstream Accuracy	22.6	58.93	**69.35**	68.23	65.61	60.22

Table 4.1: Recall and downstream accuracy for different number of facts.

The GCN consists of L hidden layers where each layer is a non-linear function $f(\cdot,\cdot)$. Specifically,

$$H^{(l)} = f(H^{(l-1)}, A) = \sigma(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(l-1)}W^{(l-1)}) \quad \forall l \in \{1,\ldots,L\}, \tag{4.1}$$

where the input to the GCN is $H^{(0)}$, $\tilde{A} = A_{\text{adj}} + I$ (I is an identity matrix), $\tilde{D}$ is the diagonal node degree matrix of $\tilde{A}$, $W^{(l)}$ is the matrix of trainable weights at the l-th layer of the GCN, and $\sigma(\cdot)$ is a non-linear activation function. We let the K-dimensional vector $\hat{g}(e) \in \mathbb{R}^K$ refer to the output of the GCN, extracted from $H^{(L)} \in \mathbb{R}^{|E|\times K}$. Hereby, K is the number of output features.

The output of the GCN, $\hat{g}(e)$ is passed through an MLP to obtain the probability $p_w^{\text{NN}}(\hat{g}(e))$ that $e \in E$ is the answer for the given question-image pair. We obtain our predicted answer $\hat{A}$ via

$$\hat{A} = \arg\max_{e\in E} p_w^{\text{NN}}(\hat{g}(e)). \tag{4.2}$$

As mentioned before, each node $e \in E$ is represented by the concatenation of the corresponding image, question and entity representation, *i.e.*, $g_w(e) = (g_w^V(I), g_w^Q(Q), g_w^C(e))$. We discuss those three representations subsequently.

1. Image Representation: The image representation, $g_w^V(I) \in \{0,1\}^{1176}$ is a multi-hot vector of size 1176, indicating the visual concepts which are grounded in the image. Three types of visual concepts are detected in the image: actions, scenes and objects. These are detected using the same pre-trained networks described in [34].

2. Question Representation: An LSTM net is used to encode each question into the representation $g_w^Q(Q) \in \mathbb{R}^{128}$. The LSTM is initialized with GloVe embeddings [**?**] for each word in the question, which is fine-tuned during training. The hidden representation of the LSTM constitutes the question encoding.

3. Entity Representation: For each question, the entity encoding $g_w^C(e)$ is computed for every entity e in the entity set E. Note that an entity e is generally composed of multiple words. Therefore, similar to the question encoding, the hidden representation of an LSTM net is used. It is also initialized with the GloVe embeddings [**?**] of each word in the entity, which is fine-tuned during training.

The answer prediction model parameters consists of weights from the question embedding, entity embedding, GCN, and MLP. These are trained end-to-end.

4.2 LEARNING

We note that the answer prediction and relation prediction model parameters are trained separately. The dataset, $\mathcal{D} = \{(I, Q, f^*, A^*)\}$, to train both these parameters is obtained from [30]. It contains tuples (I, Q, f^*, A^*) each composed of an image I, a question Q, as well as the ground-truth fact f^* and answer A^*.

To train the relation predictor's parameters we use the subset $\mathcal{D}_1 = \{(Q, r^*)\}$, containing pairs of questions and the corresponding relations $r^* = \text{rel}(f^*)$ extracted from the ground-truth fact f^*. Stochastic gradient descent and classical cross-entropy loss are used to train the classifier.

The answer predictor's parameters, consist of the question and entity embeddings, the two hidden layers of the GCN, and the layers of the MLP. The model operates on question-image pairs and extracts the entity label from the ground-truth answer A^* of the dataset $\mathcal{D}$, *i.e.*, 0 if it isn't the answer and 1 if it is. Again we use stochastic gradient descent and binary cross-entropy loss.

4.3 EXPERIMENTS

Before assessing the proposed approach subsequently, we first review properties of the FVQA dataset. We then present quantitative results to compare our proposed approach with existing baselines before illustrating qualitative results.

4.3.1 Factual visual question answering dataset

To evaluate our model, We use the publicly available FVQA [30] knowledge base and dataset. This dataset consists of 2,190 images, 5,286 questions, and 4,126 unique facts corresponding to the questions. The knowledge base consists of 193,449 facts, which were constructed by extracting top visual concepts for all images in the dataset and querying for those concepts in the knowledge bases, WebChild [31], ConceptNet [33], and DBPedia [32].

Method							Accuracy	
							@1	@3
LSTM-Question+Image+Pre-VQA [30]							24.98	40.40
Hie-Question+Image+Pre-VQA [30]							43.41	59.44
FVQA [30]							56.91	64.65
Ensemble [30]							58.76	-
Straight to the Facts (STTF) [34]							62.20	75.60
Ours	**Q**	**VC**	**Entity**	**MLP**	**GCN Layers**	**Rel**		
1	✓	-	✓	✓	-	-	10.32	13.15
2	✓	-	✓	✓	-	@1	13.89	16.40
3	✓	-	✓	✓	2	@1	14.12	17.75
4	✓	✓	✓	✓	-	-	29.72	35.38
5	✓	✓	✓	✓	-	@1	50.36	56.21
6	✓	✓	✓	-	2	@1	48.43	53.87
7	✓	✓	✓	✓	1	@1	54.60	60.91
8	✓	✓	✓	✓	1	@3	57.89	65.14
9	✓	✓	✓	✓	3	@1	56.90	62.32
10	✓	✓	✓	✓	3	@3	60.78	68.65
11	✓	✓	✓	✓	2	@1	65.80	77.32
12	✓	✓	✓	✓	2	@3	**69.35**	**80.25**
13	✓	✓	✓	✓	2	*gt*	72.97	83.01
Human							77.99	-

Table 4.2: Answer accuracy over the FVQA dataset.

4.3.2 Retrieval of Relevant Facts

As described in Sec. 4.1.1, a similarity scoring technique is used to retrieve the top-100 facts f_{100} for every question. GloVe 100d embeddings are used to represent each word in the fact and question. An initial stop-word removal is performed to remove stop words (such as "what," "where," "the") from the question. To assign a similarity score to each fact, we compute the word-wise cosine similarity of the GloVe embedding of every word in the fact with the words in the question and the detected visual concepts. We choose the top K% of the words in the fact with the highest similarity and average these values to assign a similarity score to the fact. Empirically we found $K = 80$ to give the best result. The facts are sorted based on the similarity and the 100 highest scoring facts are filtered. Tab. 4.1 shows that the ground truth fact is present in the top-100 retrieved facts 84.8% of the time and is retrieved as the top-1 fact 22.5% of the time. The numbers reported are an average over the five test sets. We also varied the number of facts retrieved in the first stage and report the recall and downstream accuracy in Tab. 4.1. The recall @50 (76.5%) is lower than

Sub-component	**Error % @1**
Fact-retrieval	15.20
Relation prediction	9.4
Answer prediction(GCN)	6.05
Total error	30.65

Table 4.3: Error contribution of the sub-components of the model to the total Top-1 error (30.65%).

the recall @100 (84.8%), which causes the final accuracy of the model to drop to 58.93%. When we retrieve 150 facts, recall is 88.4% and final accuracy is 68.23%, which is slightly below the final accuracy when retrieving 100 facts (69.35%). The final accuracy further drops as we increase the number of retrieved facts to 200 and 500.

4.3.3 Predicting the relation

As described earlier, we use the network proposed in [34] to determine the relation given a question. Using this approach, the Top-1 and Top-3 accuracy for relation prediction are 75.4% and 91.97% respectively.

4.3.4 Predicting the Correct Answer

Sec. 4.1.2 explains in detail the model used to predict an answer from the set of candidate entities E. Each node of the graph $\mathcal{G}$ is represented by the concatenation of the image, question, and entity embeddings. The image embedding $g_w^V(I)$ is a multi-hot vector of size 1176, indicating the presence of a visual concept in the image. The LSTM to compute the question embedding $g_w^Q(Q)$ is initialized with GloVe 100d embeddings for each of the words in the question. Batch normalization and a dropout of 0.5 is applied after both the embedding layer and the LSTM layer. The question embedding is given by the hidden layer of the LSTM and is of size 128. Each entity $e \in E$ is also represented by a 128 dimensional vector $g_w^C(e)$ which is computed by an LSTM operating on the words of the entity e. The concatenated vector $g_w(e) = (g_w^V(I), g_w^Q(Q), g_w^C(e))$ has a dimension of 1429 (*i.e.*, 1176+128+128).

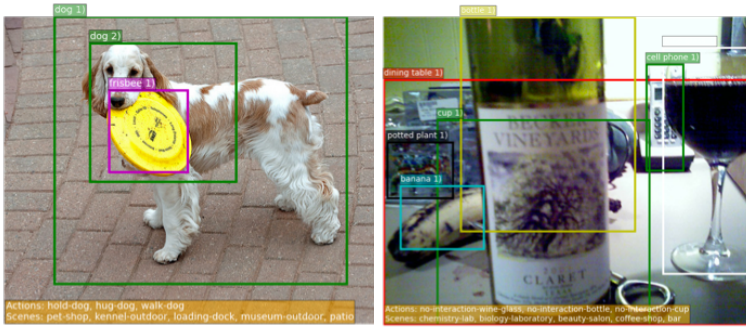
For each question, the feature matrix $H^{(0)}$ is constructed from the node representations $g_w(e)$. The adjacency matrix A_{adj} denotes the edges between the nodes. It is constructed by using the Top-1 or Top-3 relations predicted in Sec. 4.1.1. The adjacency matrix $A_{\text{adj}} \in \{0,1\}^{200\times200}$ is of size 200×200 as the set E has at most 200 unique entities (*i.e.*, 2 entities per fact and 100 facts per question). The GCN consists of 2 hidden layers, each operating on 200 nodes, and each node is represented by a feature vector of size 512. The representations of each node from the second hidden layer, *i.e.*, $H^{(2)}$ are used as input for a multi-layer perceptron which has 512 input nodes and 128 hidden nodes. The output of the hidden nodes is passed to a binary classifier that predicts 0 if the entity is not the answer and 1 if it is. The model is trained end-to-end over 100 epochs with batch gradient descent (Adam optimizer) using cross-entropy loss for each node. Batch normalization and a dropout of 0.5 was applied after each layer. The activation function used throughout is ReLU.

To prove the effectiveness of our model, we show six ablation studies in Tab. 4.2. Q, VC, Entity denote question, visual concept, and entity embeddings respectively. '11' is the model discussed in Sec. 4 where the entities are first filtered by the predicted relation and each node of the graph is represented by a concatenation of the question, visual concept, and entity embeddings. '12' uses the top three relations predicted by the question-relation LSTM net and retains all the entities which are connected by these three relations. '13' uses the ground truth relation for every question.

To validate the approach we construct some additional baselines. In '1,' each node is represented using only the question and the entity embeddings and the entities are not filtered by relation. Instead, all the entities in E are fed to the MLP. '2' additionally filters based on relation. '3' introduces a 2-layer GCN before the MLP. '4' is the same as '1' except each node is now represented using question, entity and visual concept embeddings. '5' filters by relation and skips the GCN by feeding the entity representations directly to the MLP. '6' skips the MLP and the output nodes of the GCN are directly classified using a binary classifier. We observe that there is a significant improvement in performance when we include the visual concept features in addition to question and entity embeddings, thus highlighting the importance of the visual concepts. Without visual concepts, the facts retrieved in the first step have low recall which in turn reduces the downstream test accuracy.

We also report the top-1 and top-3 accuracy obtained by varying the number of layers in the GCN. With 3 layers ('9' and '10'), our model overfits, causing the test accuracy to drop to 60.78%. With 1 layer ('7' and '8'), the accuracy is 57.89% and we hypothesize that this is due to the limited information exchange that occurs with one GCN layer. We observe a correlation between the sparsity of the adjacency matrix and the performance of the 1 layer GCN model. When the number of facts retrieved is large and the matrix is less sparse, the 1 layer GCN model makes a wrong prediction. This indicates that the 2nd layer of the GCN allows for more message passing and provides a stronger signal when there are many facts to analyze.

We compare the accuracy of our model with the FVQA baselines and our previous work,


dog 1)
dog 2)
frisbee 1)
Actions: hold-dog, hug-dog, walk-dog
Scenes: pet-shop, kennel-outdoor, loading-dock, museum-outdoor, patio
bottle 1)
dining table 1)
cell phone 1)
cup 1)
potted plant 1)
banana 1)
Actions: no-interaction-wine-glass, no-interaction-bottle, no-interaction-cup
Scenes: chemistry-lab, biology-laboratory, beauty-salon, coffee-shop, bar

Question: Which object in this image can fly?
Predicted/GT Relation: CapableOf
Relevant Visual Concept: Frisbee
Supporting Fact: (Frisbee, CapableOf, Flying)
Predicted/GT Answer: **Frisbee**

Question: Which is the yellow fruit in this image?
Predicted/GT Relation: IsA
Relevant Visual Concept: Banana
Supporting Fact: (Banana, IsA, Yellow fruit)
Predicted/GT Answer: **Banana**

Figure 4.2: Visual Concepts (VCs) detected by our model. For each image we detect objects, scenes, and actions. We observe the supporting facts to have strong alignment with the VCs which proves the effectiveness of including VCs in our model.

STTF in Tab. 4.2. The accuracy reported here is averaged over all five train-test splits. As shown, our best model '13' outperforms the state-of-the-art STTF technique by more than 7% and the FVQA baseline without ensemble by over 12%. Note that combining GCN and MLP clearly outperforms usage of only one part. FVQA and STTF both try to predict the ground truth fact. If the fact is predicted incorrectly, the answer will also be wrong, thus causing the model to fail. Our method circumvents predicting the fact and instead uses multiple relevant facts to predict the answer. This approach clearly works better.

4.3.5 Synonyms and homographs

Here we show the improvements of our model compared to the baseline with respect to synonyms and homographs. To retrieve the top 100 facts, we use trainable word embeddings which are known to group synonyms and separate homographs.

We ran additional tests using Wordnet to determine the number of question-fact pairs which contain synonyms. The test data contains 1105 such pairs out of which our model predicts 95.38% correctly, whereas the FVQA and STTF models predict 78% and the 91.6% correctly. In addition, we manually generated 100 synonymous questions by replacing words in the questions with synonyms (*e.g.* "What in the bowl can you eat?", is rephrased as, "What in the bowl is edible?"). Tests on these 100 new samples find that our model predicts 91 of these correctly, whereas the key-word matching FVQA technique gets only 61 of these right. As STTF also uses GloVe embeddings, it gets 89 correct. With regards to homographs, the test set has 998 questions which contain words that have multiple meanings across facts. Our model predicts correct answers for 81.16%, whereas the FVQA model and STTF model

Method	KB-Source Accuracy					
	DBpedia		ConceptNet		WebChild	
	Top-1	Top-3	Top-1	Top-3	Top-1	Top-3
LSTM-Question+Image+Pre-VQA [30]	15.38	32.64	25.97	41.02	34.42	50.27
Hie-Question+Image+Pre-VQA [30]	38.39	56.07	43.23	59.47	52.85	66.49
FVQA, top-1 [30]	51.25	63.07	53.50	60.16	43.54	46.58
FVQA, top-3 [30]	56.67	69.31	57.60	64.70	48.74	53.45
Ensemble [30]	57.08	-	58.98	-	59.77	-
Ours - Q_VC_Entity + GCN + MLP + Rel@1	72.97	80.76	71.20	75.82	53.33	68.69
Ours - Q_VC_Entity + GCN + MLP + Rel@3	**75.32**	**83.91**	**73.71**	**79.64**	**56.22**	**71.39**
Ours - Q_VC_Entity + GCN + MLP + *gt*-Rel	83.24	90.23	80.76	86.90	64.78	81.31
Human	74.41	-	78.32	-	81.95	-

Table 4.4: Accuracy in predicting the correct answer based on the knowledge base(KB). Best model is shown in bold.

get 66.33% and 79.4% correct, respectively.

Table 4.4 shows the top-1 and top-3 accuracy in predicting the answers on questions which contain facts from different knowledge bases. All the facts in DBpedia belong to the class “Category”. These are easy to identify as the question usually contains the term ”Category”. For example, the question, “What category of food does cake belong to?” has the supporting fact as “(Cake, Category, Herbivore)”. Our model attains an Top-1 accuracy of 75.32% on this. ConceptNet consists of the 11 relations excluding “Category” and “Comparative.” Identifying these relations from the question is pretty straightforward and we achieve a Top-1 accuracy of 73.71% in predicting the answer on questions that have facts from ConceptNet. The error in both ConceptNet and DBpedia is mainly due to some visual concepts going undetected. “Webchild” consists of facts with comparative terms (such as “faster”, “stronger”) which are also easy to identify in the given question. Our model accurately predicts the answer for these questions 56.22% of the time.

We also report the Wu-Palmer Similarity (WUPS) scores [81] in Tables 4.5 and 4.6. WUPS computes the similarity between two words based on their common subsequence in the taxonomy tree. If the similarity between the predicted and GT answer is greater than a certain threshold, the answer predicted is considered right. We report the WUPS at thresholds 0.9 and 0.0. Our model performs better than the state-of-the-art models at both thresholds.

Method	**WUPS@0.0**	
	@1	**@3**
LSTM-Question+Image+Pre-VQA [30]	63.42	76.63
FVQA, top-1 [30]	64.96	69.57
Hie-Question+Image+Pre-VQA [30]	71.51	82.71
FVQA, top-3 [30]	72.34	77.52
Ours - Q_VC_Entity + GCN + MLP + Rel@1	82.98	86
Ours - Q_VC_Entity + GCN + MLP + Rel@3	**83.55**	**88.01**
Ours - Q_VC_Entity + GCN + MLP + *gt*-Rel	85.97	90.34
Human	87.30	-

Table 4.5: WUPS@0.0 over the FVQA dataset.

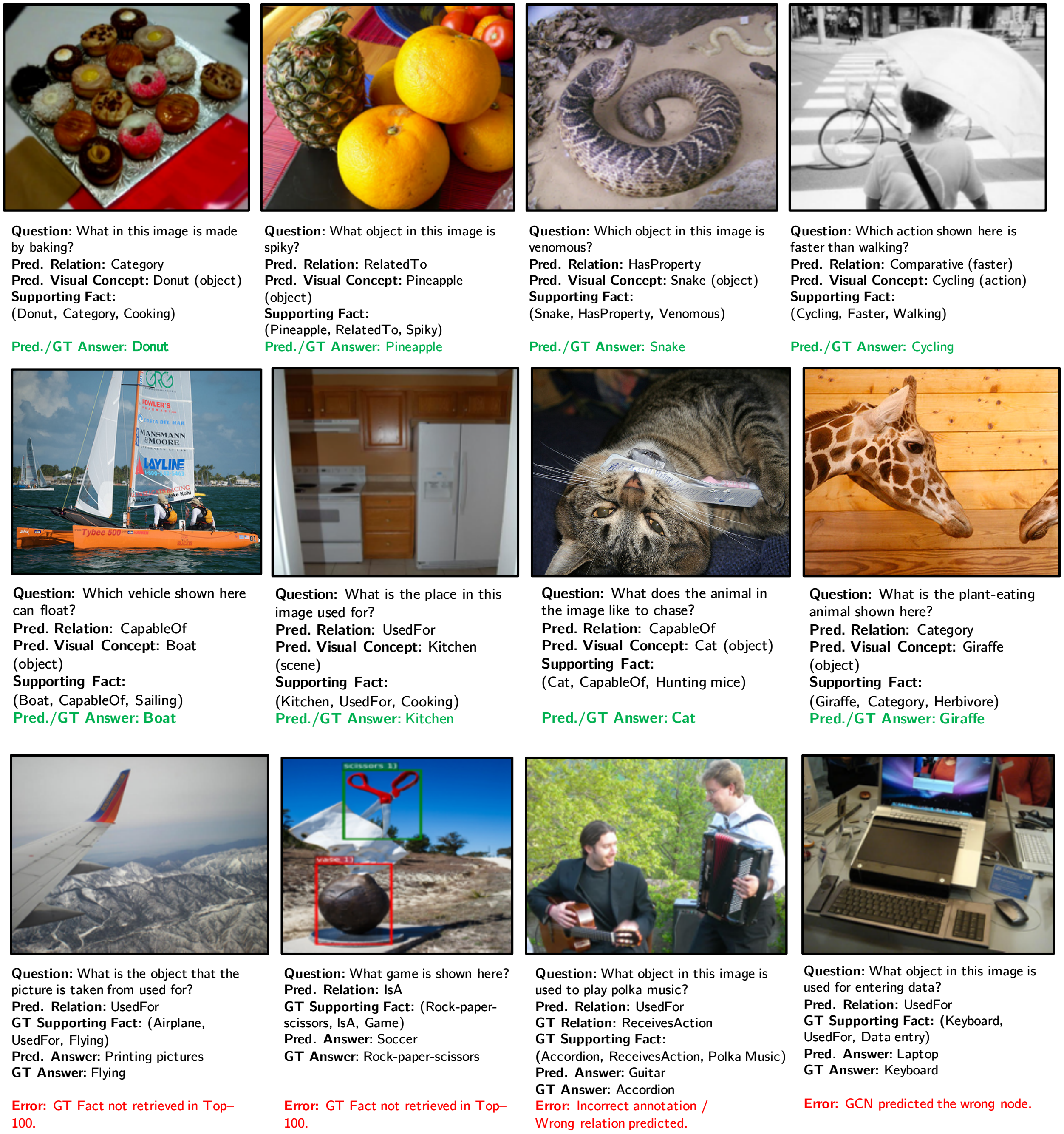
Method	**WUPS@0.9**	
	@1	**@3**
LSTM-Question+Image+Pre-VQA [30]	31.96	48.55
Hie-Question+Image+Pre-VQA [30]	48.93	64.75
FVQA, top-1 [30]	54.79	61.41
FVQA, top-3 [30]	59.67	66.89
Ours - Q_VC_Entity + GCN + MLP + Rel@1	64.70	65.12
Ours - Q_VC_Entity + GCN + MLP + Rel@3	**69.63**	**71.28**
Ours - Q_VC_Entity + GCN + MLP + *gt*-Rel	72.24	76.19
Human	82.47	-

Table 4.6: WUPS@0.9 over the FVQA dataset.

4.3.6 Qualitative results

As described in Sec. 4.1.2, the image embedding is constructed based on the visual concepts detected in the image. Fig. 4.2 shows the object, scene, and action detection for two examples in our dataset. We also indicate the question corresponding to the image, the supporting fact, relation, and answer detected by our model. Using the high-level features helps summarize the salient content in the image as the facts are closely related to the visual concepts. We observe our model to work well even when the question does not focus on the main visual concept in the image. Tab. 4.2 shows that including the visual concept improves the accuracy of our model by nearly 20%.

Fig. 4.3 depicts a few success and failure examples of our method. In our model, predicting



Question: What in this image is made by baking?
Pred. Relation: Category
Pred. Visual Concept: Donut (object)
Supporting Fact:
(Donut, Category, Cooking)

Pred./GT Answer: Donut

Question: What object in this image is spiky?
Pred. Relation: RelatedTo
Pred. Visual Concept: Pineapple (object)
Supporting Fact:
(Pineapple, RelatedTo, Spiky)
Pred./GT Answer: Pineapple

Question: Which object in this image is venomous?
Pred. Relation: HasProperty
Pred. Visual Concept: Snake (object)
Supporting Fact:
(Snake, HasProperty, Venomous)

Pred./GT Answer: Snake

Question: Which action shown here is faster than walking?
Pred. Relation: Comparative (faster)
Pred. Visual Concept: Cycling (action)
Supporting Fact:
(Cycling, Faster, Walking)

Pred./GT Answer: Cycling

Question: Which vehicle shown here can float?
Pred. Relation: CapableOf
Pred. Visual Concept: Boat (object)
Supporting Fact:
(Boat, CapableOf, Sailing)
Pred./GT Answer: Boat

Question: What is the place in this image used for?
Pred. Relation: UsedFor
Pred. Visual Concept: Kitchen (scene)
Supporting Fact:
(Kitchen, UsedFor, Cooking)
Pred./GT Answer: Kitchen

Question: What does the animal in the image like to chase?
Pred. Relation: CapableOf
Pred. Visual Concept: Cat (object)
Supporting Fact:
(Cat, CapableOf, Hunting mice)

Pred./GT Answer: Cat

Question: What is the plant-eating animal shown here?
Pred. Relation: Category
Pred. Visual Concept: Giraffe (object)
Supporting Fact:
(Giraffe, Category, Herbivore)
Pred./GT Answer: Giraffe

Question: What is the object that the picture is taken from used for?
Pred. Relation: UsedFor
GT Supporting Fact: (Airplane, UsedFor, Flying)
Pred. Answer: Printing pictures
GT Answer: Flying

Error: GT Fact not retrieved in Top–100.

Question: What game is shown here?
Pred. Relation: IsA
GT Supporting Fact: (Rock-paper-scissors, IsA, Game)
Pred. Answer: Soccer
GT Answer: Rock-paper-scissors

Error: GT Fact not retrieved in Top–100.

Question: What object in this image is used to play polka music?
Pred. Relation: UsedFor
GT Relation: ReceivesAction
GT Supporting Fact:
(Accordion, ReceivesAction, Polka Music)
Pred. Answer: Guitar
GT Answer: Accordion
Error: Incorrect annotation / Wrong relation predicted.

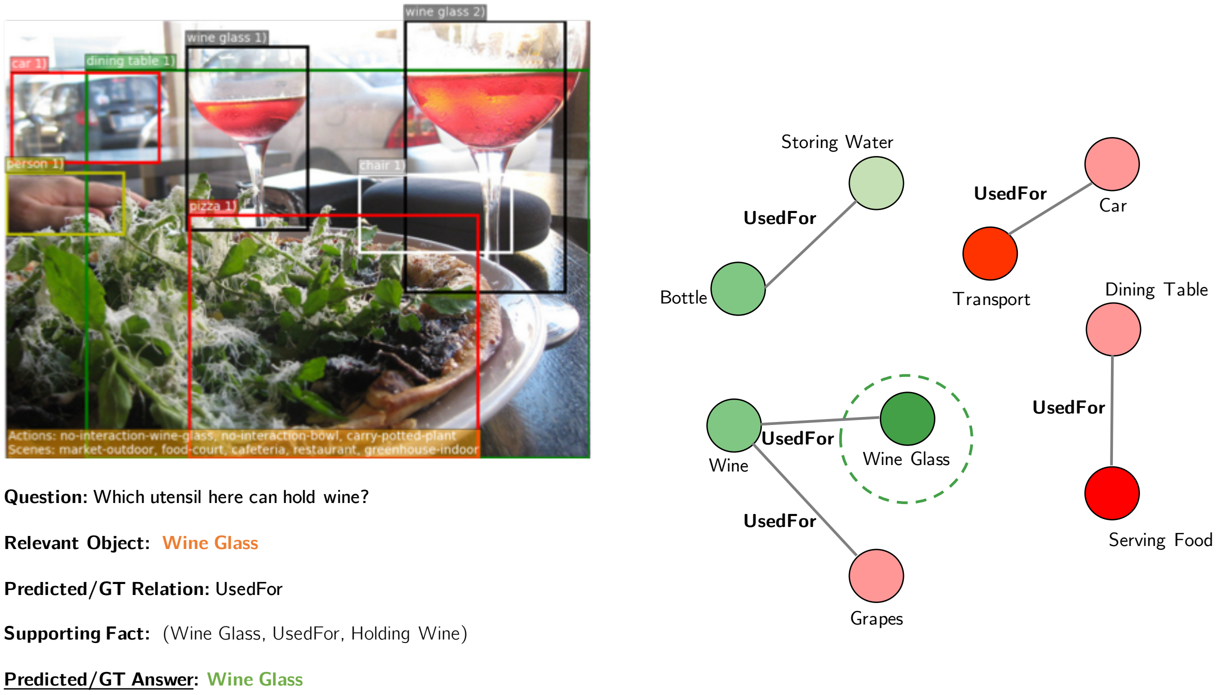
Question: What object in this image is used for entering data?
Pred. Relation: UsedFor
GT Supporting Fact: (Keyboard, UsedFor, Data entry)
Pred. Answer: Laptop
GT Answer: Keyboard

Error: GCN predicted the wrong node.

Figure 4.3: Success and failure cases: Success cases are shown in the top two rows. Our method correctly predicts the relation, visual concept, and the answer. The bottom row shows three different failure cases.

the correct answer involves three main steps: (1) Selecting the right supporting fact in the Top-100 facts, f_{100}; (2) Predicting the right relation; (3) Selecting the right entity in the GCN. In the top two rows of examples, our model correctly executes all the three steps.

As shown, our model works for visual concepts of all three types, *i.e.*, actions, scenes and objects. Examples in the second row indicates that our model works well with synonyms and homographs as we use GloVe embeddings of words. The second example in the second row shows that our method obtains the right answer even when the question and the fact do not have many words in common. This is due to the comparison with visual concepts while retrieving the facts.



Question: Which utensil here can hold wine?

Relevant Object: Wine Glass

Predicted/GT Relation: UsedFor

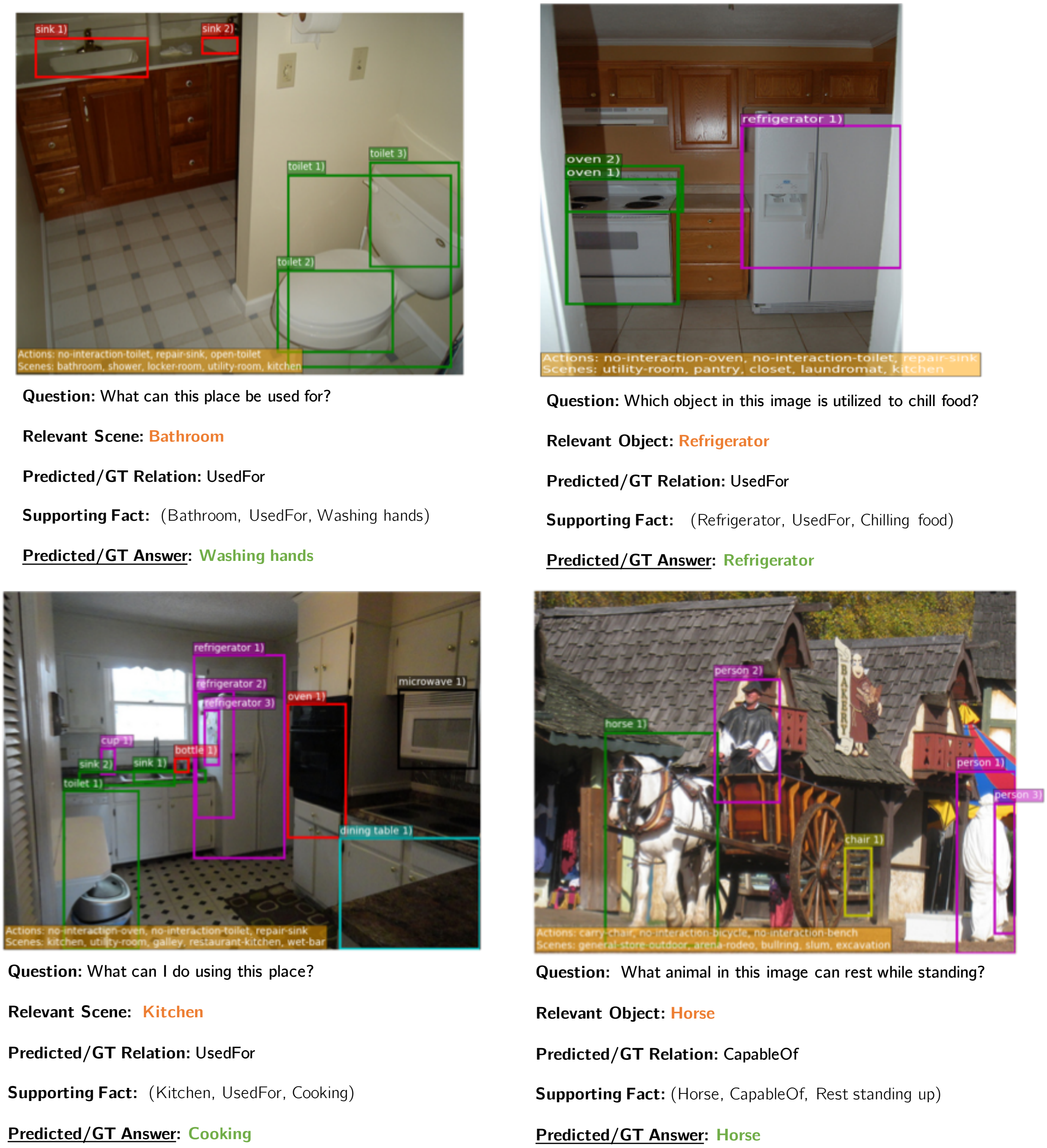
Supporting Fact: (Wine Glass, UsedFor, Holding Wine)

Predicted/GT Answer: Wine Glass

Figure 4.4: Answer prediction using GCN.

The last row shows failure cases. Our method fails if any of the three steps produce incorrect output. In the first example the ground-truth fact (Airplane, UsedFor, Flying) isn't part of the top-100. This happens when words in the fact are neither related to the words in the question nor the list of visual concepts. A second failure mode is due to wrong node/entity predictions (selecting laptop instead of keyboard), *e.g.*, because a similar fact, (Laptop, UsedFor, Data processing) exists. These type of errors are rare (Tab. 4.3) and happen only when the fact space contains a fact similar to the ground truth one. The third failure mode is due to relation prediction accuracies which are around 75%, and 92% for Top-1 and Top-3 respectively, as shown in [34].

Fig. 4.4 demonstrates the GCN's ability to reason by sharing information between the nodes and also proves the explainability of our model. The circles indicate the nodes/entities. Green indicates our model voted for this node to be an answer (*i.e.*, classified as 1). The


sink 1)
sink 2)
toilet 1)
toilet 2)
toilet 3)
Actions: no-interaction-toilet, repair-sink, open-toilet
Scenes: bathroom, shower, locker-room, utility-room, kitchen
refrigerator 1)
oven 2)
oven 1)
Actions: no-interaction-oven, no-interaction-toilet, repair-sink
Scenes: utility-room, pantry, closet, laundromat, kitchen
refrigerator 1)
refrigerator 2)
refrigerator 3)
microwave 1)
oven 1)
cup 1)
bottle 1)
sink 2)
sink 1)
toilet 1)
dining table 1)
Actions: no-interaction-oven, no-interaction-toilet, repair-sink
Scenes: kitchen, utility-room, galley, restaurant-kitchen, wet-bar
person 2)
horse 1)
person 1)
person 3)
chair 1)
Actions: carry-chair, no-interaction-bicycle, no-interaction-bench
Scenes: general-store-outdoor, arena-rodeo, bullring, slum, excavation

Question: What can this place be used for?

Relevant Scene: Bathroom

Predicted/GT Relation: UsedFor

Supporting Fact: (Bathroom, UsedFor, Washing hands)

Predicted/GT Answer: Washing hands

Question: Which object in this image is utilized to chill food?

Relevant Object: Refrigerator

Predicted/GT Relation: UsedFor

Supporting Fact: (Refrigerator, UsedFor, Chilling food)

Predicted/GT Answer: Refrigerator

Question: What can I do using this place?

Relevant Scene: Kitchen

Predicted/GT Relation: UsedFor

Supporting Fact: (Kitchen, UsedFor, Cooking)

Predicted/GT Answer: Cooking

Question: What animal in this image can rest while standing?

Relevant Object: Horse

Predicted/GT Relation: CapableOf

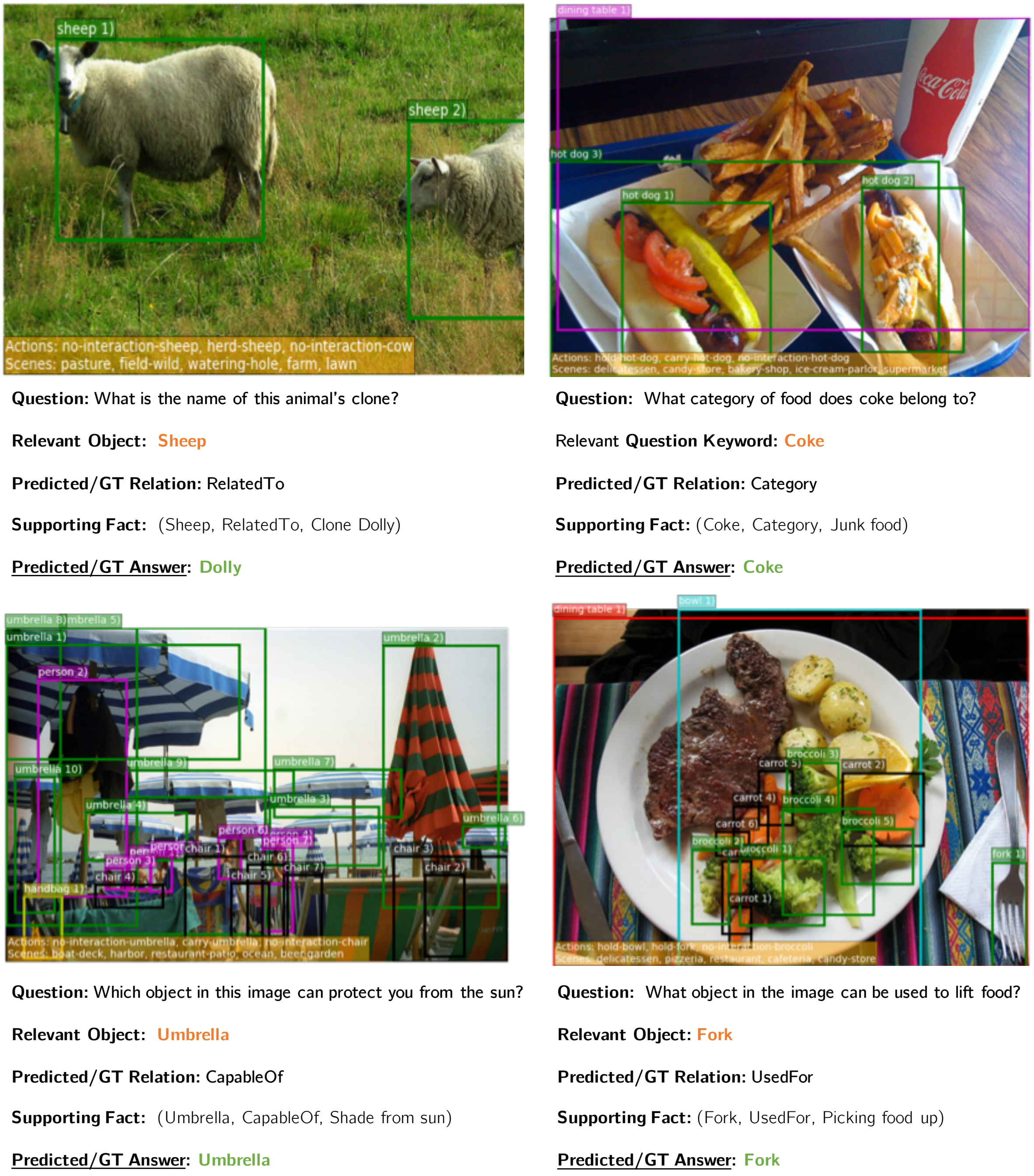
Supporting Fact: (Horse, CapableOf, Rest standing up)

Predicted/GT Answer: Horse

Figure 4.5: Success cases.

darkness corresponds to the score it was assigned. We can see that the GT node has the darkest shade. Red indicates the node was classified 0.

Fig. 4.5 shows four cases where the method proposed by [30] fails but our model works. The figures show the visual concepts detected by our predictor as well.


sheep 1)
sheep 2)
Actions: no-interaction-sheep, herd-sheep, no-interaction-cow
Scenes: pasture, field-wild, watering-hole, farm, lawn
dining table 1)
hot dog 1)
hot dog 2)
hot dog 3)
Actions: hold-hot-dog, carry-hot-dog, no-interaction-hot-dog
Scenes: delicatessen, candy-store, bakery-shop, ice-cream-parlor, supermarket
umbrella 1)
umbrella 2)
umbrella 3)
umbrella 6)
umbrella 7)
umbrella 9)
umbrella 10)
person 2)
person 3)
person 6)
person 7)
chair 1)
chair 2)
chair 3)
chair 4)
chair 5)
chair 6)
chair 7)
handbag 1)
Actions: no-interaction-umbrella, carry-umbrella, no-interaction-chair
Scenes: boat-deck, harbor, restaurant-patio, ocean, beer-garden
dining table 1)
bowl 1)
broccoli 1)
broccoli 2)
broccoli 3)
broccoli 4)
broccoli 5)
carrot 1)
carrot 2)
carrot 4)
carrot 5)
carrot 6)
fork 1)
Actions: hold-bowl, hold-fork, no-interaction-broccoli
Scenes: delicatessen, pizzeria, restaurant, cafeteria, candy-store

Question: What is the name of this animal's clone?

Relevant Object: Sheep

Predicted/GT Relation: RelatedTo

Supporting Fact: (Sheep, RelatedTo, Clone Dolly)

Predicted/GT Answer: Dolly

Question: What category of food does coke belong to?

Relevant **Question Keyword:** Coke

Predicted/GT Relation: Category

Supporting Fact: (Coke, Category, Junk food)

Predicted/GT Answer: Coke

Question: Which object in this image can protect you from the sun?

Relevant Object: Umbrella

Predicted/GT Relation: CapableOf

Supporting Fact: (Umbrella, CapableOf, Shade from sun)

Predicted/GT Answer: Umbrella

Question: What object in the image can be used to lift food?

Relevant Object: Fork

Predicted/GT Relation: UsedFor

Supporting Fact: (Fork, UsedFor, Picking food up)

Predicted/GT Answer: Fork

Figure 4.6: Success cases.

In Fig. 4.6, you can see that our model correctly predicts the object bounding boxes and the corresponding answers.

CHAPTER 5: CONCLUSIONS

In this work, we developed two novel approaches for factual visual question answering. The first approach 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. We have demonstrated the efficacy of the proposed method on the recently introduced and challenging FVQA dataset, outperforming competing methods by 5%.

The second method 'reasons' with graph convolution nets for factual visual question answering. We showed that our proposed algorithm outperforms existing baselines by a large margin of 7%. We attribute these improvements to 'joint reasoning about answers,' which facilitates sharing of information before making an informed decision. Further, we achieve this high increase in performance by using only the ground truth relation and answer information, without relying on the ground truth fact. Currently, all the components of our model except for fact retrieval are trainable end-to-end. In the future, we plan to extend our network to incorporate this step into a unified framework.

REFERENCES

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual Question Answering," in *ICCV*, 2015. 1, 4

[2] K. J. Shih, S. Singh, and D. Hoiem, "Where to look: Focus regions for visual question answering," in *CVPR*, 2016. 1, 4

[3] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *NeurIPS*, 2016. 1, 4

[4] P. Zhang, Y. Goyal, D. Summers-Stay, D. Batra, and D. Parikh, "Yin and yang: Balancing and answering binary visual questions," *arXiv:1511.05099*, 2015. 1

[5] A. Jabri, A. Joulin, and L. van der Maaten, "Revisiting Visual Question Answering Baselines," in *ECCV*, 2016. 1, 4

[6] I. Schwartz, A. G. Schwing, and T. Hazan, "High-Order Attention Models for Visual Question Answering," in *NeurIPS*, 2017. 1, 4

[7] U. Jain, Z. Zhang, and A. G. Schwing, "Creativity: Generating Diverse Questions using Variational Autoencoders," in *CVPR*, 2017. 1

[8] N. Mostafazadeh, I. Misra, J. Devlin, M. Mitchell, X. He, and L. Vanderwende, "Generating natural questions about an image," *arXiv:1603.06059*, 2016. 1

[9] Y. Li, N. Duan, B. Zhou, X. Chu, W. Ouyang, and X. Wang, "Visual question generation as dual task of visual question answering," *arXiv:1709.07192*, 2017. 1

[10] A. K. Vijayakumar, M. Cogswell, R. R. Selvaraju, Q. Sun, S. Lee, D. Crandall, and D. Batra, "Diverse beam search: Decoding diverse solutions from neural sequence models," *arXiv:1610.02424*, 2016. 1

[11] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra, "Visual Dialog," in *CVPR*, 2017. 1

[12] A. Das, S. Kottur, J. M. Moura, S. Lee, and D. Batra, "Learning cooperative visual dialog agents with deep reinforcement learning," *arXiv:1703.06585*, 2017. 1

[13] U. Jain, S. Lazebnik, and A. G. Schwing, "Two can play this Game: Visual Dialog with Discriminative Question Generation and Answering," in *CVPR*, 2018. 1

[14] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, "Visual7W: Grounded Question Answering in Images," in *CVPR*, 2016. 1, 4

[15] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A neural-based approach to answering questions about images," in *ICCV*, 2015. 1, 4

[16] M. Ren, R. Kiros, and R. Zemel, "Exploring models and data for image question answering," in *NeurIPS*, 2015. 1, 4

[17] C. Xiong, S. Merity, and R. Socher, "Dynamic memory networks for visual and textual question answering," in *ICML*, 2016. 1, 4

[18] H. Xu and K. Saenko, "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering," in *ECCV*, 2016. 1, 4

[19] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *CVPR*, 2016. 1, 4

[20] 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," in *EMNLP*, 2016. 1, 4

[21] J.-H. Kim, S.-W. L. D.-H. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, and B.-T. Zhang, "Multimodal residual learning for visual qa," in *NeurIPS*, 2016. 1, 4

[22] C. L. Zitnick, A. Agrawal, S. Antol, M. Mitchell, D. Batra, and D. Parikh, "Measuring machine intelligence through visual question answering," *AI Magazine*, 2016. 1, 4

[23] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Deep compositional question answering with neural module networks," in *CVPR*, 2016. 1, 4

[24] 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?" in *EMNLP*, 2016. 1, 4

[25] L. Yu, E. Park, A. Berg, and T. Berg, "Visual madlibs: Fill in the blank image generation and question answering," in *ICCV*, 2015. 1, 4

[26] B. Zhou, Y. Tian, S. Sukhbataar, A. Szlam, and R. Fergus, "Simple baseline for visual question answering," *arXiv:1512.02167*, 2015. 1, 4

[27] Q. Wu, C. Shen, A. van den Hengel, P. Wang, and A. Dick, "Image captioning and visual question answering based on attributes and their related external knowledge," *arXiv:1603.02814*, 2016. 1, 4

[28] L. Ma, Z. Lu, and H. Li, "Learning to answer questions from image using convolutional neural network," in *AAAI*, 2016. 1, 4

[29] D. Gordon, A. Kembhavi, M. Rastegari, J. Redmon, D. Fox, and A. Farhadi, "Iqa: Visual question answering in interactive environments," in *CVPR*, 2018. 1

[30] P. Wang, Q. Wu, C. Shen, A. Dick, and A. v. d. Hengel, "Fvqa: Fact-based visual question answering," *TPAMI*, 2018. 1, 3, 4, 5, 11, 12, 13, 15, 16, 17, 18, 19, 20, 27, 28, 29, 31, 33, 34, 38, 39, 42

[31] N. Tandon, G. de Melo, F. Suchanek, and G. Weikum, "Webchild: Harvesting and organizing commonsense knowledge from the web," in *WSDM*, 2014. 1, 5, 12, 33

[32] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: A nucleus for a web of open data," in *ISWC/ASWC*, 2007. 1, 4, 5, 13, 33

[33] R. Speer, J. Chin, and C. Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge." in *AAAI*, 2017. 1, 5, 12, 33

[34] M. Narasimhan and A. G. Schwing, "Straight to the facts: Learning knowledge base retrieval for factual visual question answering," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 451–468. 2, 3, 4, 5, 31, 32, 34, 35, 41

[35] M. Narasimhan, S. Lazebnik, and A. Schwing, "Out of the box: Reasoning with graph convolution nets for factual visual question answering," in *Advances in Neural Information Processing Systems*, 2018. 2, 4

[36] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv:1609.02907*, 2016. 2, 5

[37] M. Malinowski and M. Fritz, "A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input," in *NeurIPS*, 2014. 4

[38] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu, "Are you talking to a machine? Dataset and Methods for Multilingual Image Question Answering," in *NeurIPS*, 2015. 4

[39] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning," in *CVPR*, 2017. 4

[40] H. Ben-younes, R. Cadene, M. Cord, and N. Thome, "Mutan: Multimodal tucker fusion for visual question answering," in *ICCV*, 2017. 4

[41] A. Jiang, F. Wang, F. Porikli, and Y. Li, "Compositional memory for visual question answering," *arXiv:1511.05676*, 2015. 4

[42] L. S. Zettlemoyer and M. Collins, "Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars," in *UAI*, 2005. 4

[43] L. S. Zettlemoyer and M. Collins, "Learning context-dependent mappings from sentences to logical form," in *ACL*, 2005. 4

[44] J. Berant, A. Chou, R. Frostig, and P. Liang, "Semantic Parsing on Freebase from Question-Answer Pairs," in *EMNLP*, 2013. 4

[45] Q. Cai and A. Yates, "Large-scale Semantic Parsing via Schema Matching and Lexicon Extension," in *ACL*, 2013. 4

[46] P. Liang, M. I. Jordan, and D. Klein, "Learning dependency-based compositional semantics," in *Computational Linguistics*, 2013. 4

[47] T. Kwiatkowski, E. Choi, Y. Artzi, and L. Zettlemoyer, "Scaling semantic parsers with on-the-fly ontology matching," in *EMNLP*, 2013. 4

[48] J. Berant and P. Liang, "Semantic parsing via paraphrasing," in *ACL*, 2014. 4

[49] A. Fader, L. Zettlemoyer, and O. Etzioni, "Open question answering over curated and extracted knowledge bases," in *KDD*, 2014. 4

[50] S. W. t. Yih, M.-W. Chang, X. He, and J. Gao, "Semantic parsing via staged query graph generation: Question answering with knowledge base," in *ACL-IJCNLP*, 2015. 4

[51] S. Reddy, O. Täckström, M. Collins, T. Kwiatkowski, D. Das, M. Steedman, and M. Lapata, "Transforming dependency structures to logical forms for semantic parsing," in *ACL*, 2016. 4

[52] C. Xiao, M. Dymetman, and C. Gardent, "Sequence-based structured prediction for semantic parsing," in *ACL*, 2016. 4

[53] Y. Zhang, K. Liu, S. He, G. Ji, Z. Liu, H. Wu, and J. Zhao, "Question answering over knowledge base with neural attention combining global knowledge information," *arXiv:1606.00979*, 2016. 4

[54] C. Unger, L. Bühmann, J. Lehmann, A.-C. N. Ngomo, D. Gerber, and P. Cimiano, "Template-based question answering over RDF data," in *WWW*, 2012. 4

[55] O. Kolomiyets and M.-F. Moens, "A survey on question answering technology from an information retrieval perspective," in *Information Sciences*, 2011. 4

[56] X.Yao and B. V. Durme, "Information extraction over structured data: Question answering with Freebase," in *ACL*, 2014. 4

[57] A. Bordes, S. Chopra, and J. Weston, "Question answering with sub-graph embeddings," in *EMNLP*, 2014. 4

[58] A. Bordes, J. Weston, and N. Usunier, "Open question answering with weakly supervised embedding models," in *ECML*, 2014. 4

[59] L. Dong, F. Wei, M. Zhou, and K. Xu, "Question answering over freebase with multi-column convolutional neural networks," in *ACL*, 2015. 4

[60] A. Bordes, N. Usunier, S. Chopra, and J. Weston, "Large-scale simple question answering with memory networks," in *ICLR*, 2015. 4

[61] Y. Zhu, C. Zhang, C. Ré, and L. Fei-Fei, "Building a large-scale multimodal Knowledge Base for Visual Question Answering," in *CoRR*, 2015. 4

[62] Q. Wu, P. Wang, C. Shen, A. Dick, and A. van den Hengel, "Ask me anything: Free-form visual question answering based on knowledge from external sources," in *CVPR*, 2016. 4

[63] P. Wang, Q. Wu, C. Shen, A. Dick, and A. Van Den Henge, "Explicit knowledge-based reasoning for visual question answering," in *IJCAI*, 2017. 4

[64] K. Narasimhan, A. Yala, and R. Barzilay, "Improving information extraction by acquiring external evidence with reinforcement learning," in *EMNLP*, 2016. 4

[65] J. Krishnamurthy and T. Kollar, "Jointly learning to parse and perceive: Connecting natural language to the physical world," in *ACL*, 2013. 4

[66] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *IEEE*, 1998. 5

[67] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. v. d. Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," in *ESWC*, 2018. 5

[68] X. Wang, Y. Ye, and A. Gupta, "Zero-shot recognition via semantic embeddings and knowledge graphs," in *CVPR*, 2018. 5

[69] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *EMNLP*, 2014. [Online]. Available: http://www.aclweb.org/anthology/D14-1162 9

[70] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, 1997. 9

[71] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016. 9

[72] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009. 9

[73] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NeurIPS*, 2015. 9

[74] 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. 9

[75] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *IJCV*, 2015. 9

[76] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *TPAMI*, 2017. 10

[77] A. Mallya and S. Lazebnik, "Learning models for actions and person-object interactions with transfer to question answering," in *ECCV*, 2016. 10

[78] Y.-W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng, "Hico: A benchmark for recognizing human-object interactions in images," in *ICCV*, 2015. 10

[79] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *CVPR*, 2014. 10

[80] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large Margin Methods for Structured and Interdependent Output Variables," *JMLR*, 2005. 12

[81] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *ACL*, 1994. 19, 38

APPENDIX A: ORIGINAL AND SYNONYMOUS QUESTIONS

Original: Which object in this image is related to wool? **Synonymous:** Which object in this image is related to fleece?

Original: Which object on the table is related to pizzeria? **Synonymous:** Which object on the table is related to bakery?

Original: Name the mammal in the image. **Synonymous:** Name the animal in the image.

Original: Which object in this image can be found at the corner of two streets? **Synonymous:** Which object in this image can be found at an intersection?

Original: Which object in this image is better than chopstick? **Synonymous:** What shown here is preferred over chopsticks?

Original: What are you likely to find in zoos? **Synonymous:** What are you likely to find in sanctuaries?

Original: Which object in this image is sour? **Synonymous:** Which object shown in this picture is tart?

Original: Which animal in the image can jump over hurdles? **Synonymous:** Which animal in the image can leap over hurdles?

Original: Which object in this image is related to horse? **Synonymous:** Which object in this image is related to pony?

Original: What is the dog lying on ? **Synonymous:** What is the canine lying on ?

Original: Which object in this image is related to carry on? **Synonymous:** Which object in this image is related to transport?

Original: Which object in this image is used to carry your things while travelling? **Synonymous:** Which object in this image used to carry your things while moving?

Original: Which animal in this image is less small than goat? **Synonymous:** Which animal in this picture is bigger than a goat?

Original: What does a hotdog have? **Synonymous:** What does a hotdog contain?

Original: Which object in this image is a popular pet? **Synonymous:** object in this picture is a favored pet?

Original: which object in this image is more efficient than bus? **Synonymous:** Which object in this image is more effective than a bus?

Original: Which fruit in this image is spiky **Synonymous:** Which fruit in this image is pointed?

Original: What object can be used to travel through many time zones? **Synonymous:**

What object can be used to journey through many time zones?

Original: Which fruit in this image can be found din a monkey's hand? **Synonymous:** Which fruit in this image is a monkey most likely to hold?

Original: which object in this image is a means of transportation? **Synonymous:** Which object in this image is a means of moving?

Original: Is this office neat and tidy? **Synonymous:** Is this office clean?

Original: Which object in this image is used for going down a river? **Synonymous:** Which object in this image is used for transporting down a river?

Original: Which object in this image is used for freezing food? **Synonymous:** Which object in this image is used to frost food?

Original: Which object in this image is used to measure liquid? **Synonymous:** Which object in this image is used to measure fluids?

Original: Which objects in this image are considered as food? **Synonymous:** What is the food shown in this picture?

Original: Which object in this image is a vehicle? **Synonymous:** Which object shown here is a means of transport?

Original: what object in this image contains chlorophyl? **Synonymous:** Which object shown in this picture has chlorophyl?

Original: Which object in this image is used for playing songs? **Synonymous:** Which object in this image is used for playing music?

Original: Which container is shown in this image ? **Synonymous:** Which vessel in shown in this image?

Original: What can you do in this place? **Synonymous:** What can you do in this location?

Original: Which object in this image is formal clothing? **Synonymous:** What object in this image is formal attire?

Original: Which object in the image is large? **Synonymous:** Which object in the image is huge?

Original: What in this image could be used for holding cereal? **Synonymous:** What in this image could be used for pouring cereal?

Original: Which thing in this picture is used for sailing? **Synonymous:** Which thing in this picture is used for boating?

Original: Which object in this image has stinky breath? **Synonymous:** Which object in this image has smelly breath?

Original: Which fast food in this image contains meat? **Synonymous:** Which fast food in this image contains animal flesh?

Original: What sort of food can you see in this image? **Synonymous:** What sort of eatables can you see in this image?

Original: where is this place? **Synonymous:** Where is this location?

Original: What object in this image is used for signaling danger? **Synonymous:** What object in this image is used for communicating danger?

Original: Which object in this image eats mice? **Synonymous:** Which object in this image eats rats?

Original: Which thing in the image can be used to hold water? **Synonymous:** Which thing in the image can be used to store water?

Original: Which object in this image is an aircraft? **Synonymous:** Which object in this image is a airplane?

Original: Which instrument in this image has strings? **Synonymous:** Which object in this image has strings?

Original: What animal in the image can follow its master? **Synonymous:** What animal in this image can follow its owner?

Original: Which object in this image can hold liquid? **Synonymous:** Which object in this image can hold fluids?

Original: What object in the image can be used for cuddling? **Synonymous:** what object in the image can be used for hugging?

Original: what thing is expensive than the object shown in the middle of this image? **Synonymous:** What thing is more costly than the object shown in the middle of this image?

Original: What in this image could be used for writing email? **Synonymous:** What in this image can be used for sending email?

Original: what kind of sports are people doing in this image? **Synonymous:** what kind of activity are people doing in this image?

Original: Which object in this image has fur? **Synonymous:** Which object in this image has fleece?

Original: Which object in this image could potentially drive a lorry? **Synonymous:** Which object in this image could potentially drive a truck?

Original: Which animal in this image can build nest? **Synonymous:** Which animal in this image can construct nest?

Original: What container is visible in this image? **Synonymous:** What vessel is visible in this image?

Original: what object in this image is used for cycling? **Synonymous:** What object in this image is used for biking?

Original: which object in this image could run faster than pedestrian **Synonymous:** Which object in this image could run faster than person?

Original: Which object in this image is related to SMS? **Synonymous:** Which object is this image is related to message?

Original: Which object in this image could be used to make glass? **Synonymous:** object here can be used to manufacture glass?

Original: What thing in this image can scratch? **Synonymous:** What in this image can scrape?

Original: What is the herbivore in this image? **Synonymous:** What is the plant-eating animal in this image?

Original: which object in this image can clean a messy room? **Synonymous:** Which object in the image can clean a dirty room?

Original: Which object in this image is used in a bloody mary? **Synonymous:** Which object in this image is used in a drink?

Original: which object in this image costs less money than taxi? **Synonymous:** Which object in this image is cheaper than a cab?

Original: Which object in this image might be used to make cake? **Synonymous:** Which object in this image might be used to bake cake?

Original: what object in this image can cry? **Synonymous:** What object in this image can weep?

Original: Which spectator sport is shown in this image? **Synonymous:** Which spectator game is shown in this image?

Original: What can I do in this place? **Synonymous:** What can I do in this site?

Original: Which object in this image might occur in a nightmare? **Synonymous:** Which object in this image might occur in a dream?

Original: Which part of the animal can hurt people? **Synonymous:** Which part of this animal can harm people?

Original: What can be found in this place? **Synonymous:** What can be found in this location?

Original: Which object in this image is soft and fuzzy? **Synonymous:** Which object in this image is soft and woolly?

Original: Which skiing equipment can you see in this image? **Synonymous:** Which skiing tool can you see in this image?

Original: Which object in this image has feathers? **Synonymous:** Which object in this image has plume?

Original: Which animal in this image is very big? **Synonymous:** Which animal in this image is very large?

Original: This food is normally served in what events? **Synonymous:** This food is normally served in which occasions?

Original: What's in the bowl that you can eat? **Synonymous:** What in the bowl is edible?

Original: which object in this image can chase a ball? **Synonymous:** Which object in this image can run after/fetch a ball?

Original: Which object in this image has the property of poisonous? **Synonymous:** Which object in this image is toxic?

Original: Which animal in this image can build nest? **Synonymous:** animal in this image can construct a nest?

Original: Which object in this image is able to cut food? **Synonymous:** Which object in this picture can chop food?

Original: Which object in this image is used to keep people dry when it rains? **Synonymous:** Which object in this image is used for shielding rainfall?

Original: Which object in this image is used to assemble food? **Synonymous:** Which object in this picture can be used for arranging food?

Original: Which object will you choose to cut a potato? **Synonymous:** Which object here can slice a potato?

Original: Which object in this image belongs to greenery? **Synonymous:** Which object in this image belongs to verdure?

Original: Which object in this image is used for cooling things? **Synonymous:** Which object in this picture is used for keeping things chill?

Original: What object in this image is a type of sports equipment? **Synonymous:** What object in the picture is a kind of sports gear?

Original: What instrument can be found in the living room? **Synonymous:** What tool can be found in the living room?

Original: Which object in this image is round? **Synonymous:** Which object shown here is spherical?

Original: which object in this image is used for transportation? **Synonymous:** Which object shown here is used for conveyance?

Original: which object in this image has sharp claws? **Synonymous:** Which object here has razor-edged claws?

Original: Which object in this image is capable of keeping ice cold? **Synonymous:** Which object in this picture can cool ice?

Original: Which object in this image is used for fun? **Synonymous:** Which object shown here is used for entertainment?

Original: Which object in this image are dangerous? **Synonymous:** What object in this image is not safe?

Original: What kind of ball sport is this in this image? **Synonymous:** What kind of ball game is shown here?

Original: Which object in this image has feelings? **Synonymous:** Which object in this image has emotions?

Original: What sport is a kind of good exercise? **Synonymous:** Which game is a kind of good workout?