Understanding Knowledge Gaps in Visual Question Answering: Implications for Gap Identification and Testing

Goonmeet Bajaj [*] Bortik Bandyopadhyay [*] Daniel Schmidt [†] Pranav Maneriker [*]
Christopher Myers [‡] Srinivasan Parthasarathy [*]
{bajaj.32,bandyopadhyay.14,maneriker.1}@osu.edu, schmidt.152@wright.edu,
christopher.myers.29@us.af.mil, srini@cse.ohio-state.edu

Abstract

Traditional Visual Question Answering (VQA) datasets typically contain questions related to the spatial information of objects, object attributes, or general scene questions. Recently, researchers have recognized the need to improve the balance of such datasets to reduce the system's dependency on memorized linguistic features and statistical biases, while aiming for enhanced visual understanding. However, it is unclear whether any latent patterns exist to quantify and explain these failures. As an initial step towards better quantifying our understanding of the performance of VQA models, we use a taxonomy of Knowledge Gaps (KGs) to tag questions with one or more types of KGs. Each KG describes the reasoning abilities needed to arrive at a resolution, and failure to resolve gaps indicates an absence of the required reasoning ability. After identifying KGs for each question, we examine the skew in the distribution of questions for each KG. We then introduce a targeted question generation model to reduce this skew, which allows us to generate new types of questions for an image.

1. Introduction

When compared to artificially intelligent (AI) systems, human cognition demonstrates a reasonably flexible system when faced with gaps in knowledge while executing a prescribed task. Humans often demonstrate both the ability to identify gap(s) in their knowledge and the ability to resolve these different gaps through diverse strategies (e.g., by seeking clarification, conducting research, *etc.*).

Informally, a knowledge gap (KG) is an instance of limited or missing information or capabilities, which leads to an AI agent being inefficient or incapable of completing a given task. AI agents, when presented a same/similar task as humans, might not always have the perfect knowledge to complete it [8]. A framework for KG identification and resolution can facilitate flexibility for an AI agent during both training and execution. As a preliminary effort to understand how to build a system for AI agents to *detect*, *identify*, and *resolve* the KGs that can occur, we leverage *visual question answering* (VQA) tasks. As a first step, we aim to detect gaps in knowledge and identify the knowledge gap type in this work. VQA sits at an intersection of three components of artificial intelligence: *language*, *vision*, and *reasoning*, thus making it a challenging task. To the best of our knowledge, we are the first to systematically manipulate VQA questions to produce knowledge gaps within AI agents.

Using a refined version of a KG taxonomy [1], we identify eight different KGs that occur in the GQA dataset [4]. Figure 1 shows the skew we observe in the distribution of the number of questions per KG category. To alleviate this skew and make questions more evenly distributed across KGs, we apply a neural framework to generate questions for specific KGs.

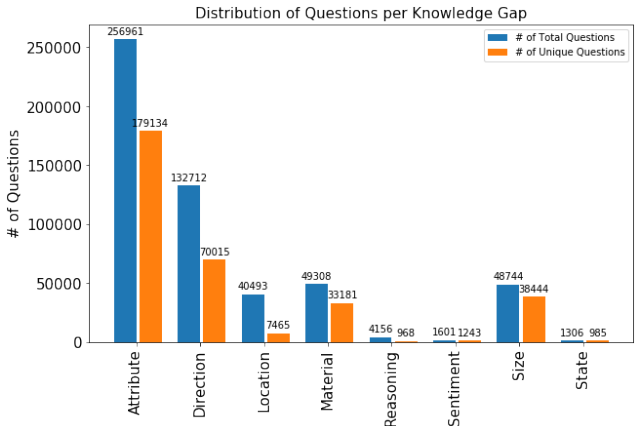

Figure 1: Skew in the Distribution of Questions per KG

[*]The Ohio State University (OSU)
[†]Wright State University
[‡]Air Force Research Laboratory (AFRL)

2. Related work

Commonsense reasoning for VQA agents: Recent VQA datasets are a result of the realization that VQA agents can be highly contingent upon statistical biases and tendencies of the answer distribution and linguistic features [2, 9]. Hudson and Manning [4] create the GQA dataset to gain control over the answer distribution and mitigate the heavy dependence on linguistic features and priors in current VQA frameworks. The FVQA [7], OK-VQA [6], and CRIC [3] datasets all contain questions in which the image content is not sufficient to answer the questions. These questions typically require external information or commonsense reasoning to answer them and result in degraded performance of state-of-the-art VQA models. To the best of our knowledge, we are the first to tag VQA questions with the reasoning skills required to answer them. Our approach provides a new channel to analyze the performance of VQA agents using different KG categories.

Question generation: To overcome the skew in the distribution of questions per KG, we generate question templates and populate them with image annotations to create new questions. Liu *et al*. [5] generate question and answer pairs using a neural network architecture to transduce knowledge base facts into natural language questions. They combine template-based question generation techniques and sequence-to-sequence learning approaches to generate new questions. We closely follow their approach to create questions that complement the GQA dataset.

3. Definition of knowledge gaps

There are five major types of KGs in the taxonomy presented in Figure 2, namely: Language, Spatial, Attribute, Reasoning, and Philosophical Gaps. **Language gaps** arise when unknown phrases or vocabulary words are introduced. **Spatial gaps** occur when there is an error in understanding the physical space of a given setting. **Attribute gaps** can occur when an object's (or person's) characteristics are not well understood. **Reasoning gaps** indicate that an agent has difficulty in the cognitive process of understanding information. **Philosophical gaps** are similar to reasoning gaps but require meta-cognitive processes.

We focus on eight KGs identified in the GQA dataset (colored in Figure 2): Attribute, Direction, Location, Material, Reasoning, Sentiment, Size, and State Gaps. We now define and ground these specific KGs for the VQA setting. We refer the readers to [1] for an extended version of the KG taxonomy and definitions.

Location Gap: Location gaps can occur when there is a misunderstanding about a specific physical place or setting of a context. We assign these gaps to questions in the dataset about the location of a scene of an image.

Reasoning Gap: We tag questions that require external knowledge about the scene or objects in an image with reasoning gaps.

Sentiment Gap: A sentiment gap can occur when an agent is not able to understand the emotion or attitude of another agent. Questions about the sentiment of an object (typically humans or animals) are marked with this gap.

Size Gap: Size gaps (subtype of attribute gap), can occur when an agent is trying to understand the physical space an object or a person occupies. Questions inquiring about the size, age, or height of an object are marked with size gap.

We use these KG definitions to create a rule-based tagging system to automatically mark questions with their respective KGs. Below, we present sample questions and the KGs assigned by our system:

- *Which are less healthy, the brownies or the cherries?*
  KG: Reasoning
- *What is the device that the happy man is holding?*
  KG: Sentiment
- *Is the large propeller blue and still?*
  KGs: Attribute, Size, State
- *Where is the horse that looks white and brown walking?*
  KGs: Attribute, Location

Due to space restrictions, we choose to focus on the following KGs: Location, Reasoning, Sentiment, and Size Gaps. The extended version of this paper [1] contains details for all eight KGs identified in the GQA dataset.

4. Knowledge gap identification

The GQA dataset [4] consists of 22M questions about various day-to-day images. We choose to work with the training dataset and direct the readers to the official GQA website[1] for more information about the original dataset. We use the following annotations associated with each question to automatically identify KGs: *detailed type*, *global group*, and *semantic filters* (see [4, 1] for more details). For example, if a question has "*placeVerify*" as its detailed type, we tag the question with a *location gap*.

First, we assign a KG based on the *detailed type* of a question. Next, we examine a question's *global group* to allot a KG, which has not been previously assigned by the *detailed type*. Lastly, we attempt to designate a KG using the *semantic filters* extracted from a question's functional program. During each step of the pipeline, we try to assign one KG and do not reassign previously specified KGs. Now, we define the mapping of dataset annotations for each KG (see [1] for a complete mapping of all eight KGs).

Location - *detailed types*: place, placeVerify, placeVerifyC, placeChoose, locationVerifyC, locationVerify; *global group*: place, room, nature environment, urban environment, road; *semantic filters*: location, place, room

[1] https://cs.stanford.edu/people/dorarad/gqa/about.html

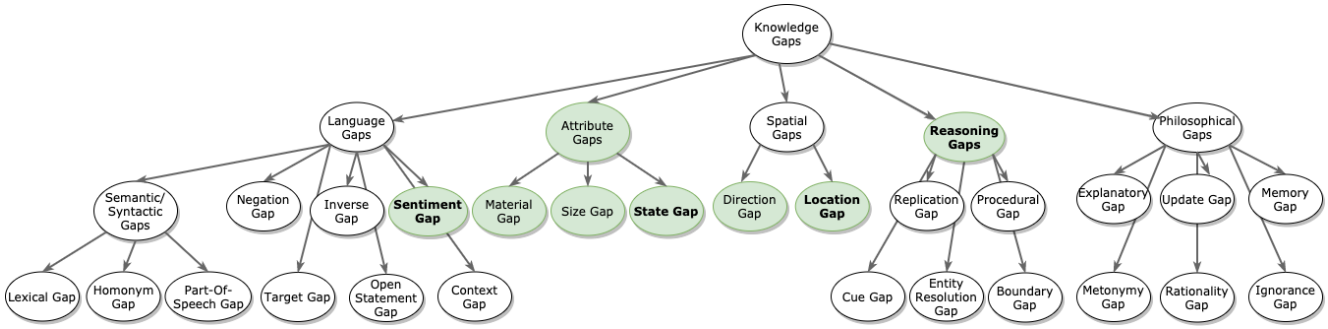

Figure 2: KG Taxonomy [1] - KGs identified in the GQA dataset are colored, and we focus on the KGs in **bold font**.

Reasoning - *detailed types*: diffAnimals, diffAnimalsC, sameAnimals, sameAnimalsC, comparativeChoose; *global group*: None; *semantic filters*: None
Sentiment - *detailed types*: None; *global group*: face expression; *semantic filters*: face expression
Size - *detailed types*: None; *global group*: age, height, thickness, depth, fatness, length, weight, width, size; *semantic filters*: age, fatness, length, thickness, size, weight, depth, width, height

As mentioned, we observe a skew in the distribution of questions for each KG category (see Figure 1). Less than 3% of the questions are tagged with reasoning, sentiment, and state gaps. About 7% of the questions are allotted with a location gap. Around 9% of the questions are tagged with a material gap and another 9% with a size gap. Attribute gaps and direction gaps are assigned to about 48% and 24% of the questions. This apparent skew inspires us to use a question generation technique to balance dataset (in terms of KGs). The extended version of the paper contains a detailed description of the KG identification methodology [1].

5. Question generation

We aim to generate new questions for images that lack certain KGs to remove the skew in the distribution of questions for each KG. We use the IBM Pytorch Seq2Seq framework[2] to train our models.

The template-based seq2seq (consisting of an encoder and decoder) model can be viewed as a translator that converts structured data (paths along a scene graph) into question templates. These templates are populated in a downstream task with information from the scene graph to generate a complete question. **For example, given a training path and question pair,** *"fries to the right of lettuce"*, **we first transform the original question,** *"Which is less healthy, the fries or the lettuce?"*, **into a question template,** *"Which is less healthy, the OBJ or the OBJ ?"*, **and use it for training.** Our neural question generation strategy resembles that of [5] and is modelled as a probabilistic framework.

$$P(Q|P) = \prod_{i=1}^{N} P(w_i|w_{<i}, P) \quad (1)$$

$Q = (w_1, w_2, ..., w_n)$ represents a generated question template that consists of tokens $w_1, w_2, ..., w_n$.

P represents a path sequence of length L that is fed into the seq2seq model's encoder. $P = (g(o_1), r_1, IO, r_2, IO, ..., r_n, g(o_2))$, where o_1 and o_2 are objects mentioned in the training questions. The function $g(\cdot)$ describes objects as a concatenation of their attributes and name in English. The r_n(s) along a path are the relations among objects. We replace the intermediate objects along a path with "IO" because we are only interested in objects o_1 and o_2 that are present in the training question.
Encoder: The (Bi)LSTM encoder encodes a path sequence, P, from a scene graph (of an image) into an embedding.
Template generation: To create training question templates, we replace the objects and attributes mentioned in the question with "OBJ" and "ATTRIBUTE" placeholders.
Decoder: We use the built-in attention mechanism with an LSTM decoder to make use of the alignment information between scene graph paths and question templates. During decoding, we use teacher forcing (ratio = 0.25) to allow the decoder to learn how to generate question templates. We use the *TopKDecoder*, which performs a beam search of length $K = 10$. Of the top decoded sequences, we select the template with the highest probability and with the same number of "ATTRIBUTE" placeholders in the generated template as the original training template. If none of the top K results meet this criterion, we use the template with the highest probability as our output.

5.1. Experimental setup

For each KG type, we train two seq2seq models: 1) *triple-based model* ($L = 1$), 2) *path-based model* ($L \leq 5$).

[2] https://ibm.github.io/pytorch-seq2seq/public/index.html

KG	BLEU Score	Meteor Score	# of Novel Templates	# of Existing Templates
Triple-based Model Results				
Location	0.32	0.64	35	14
Reasoning	0.75	0.87	7	5
Sentiment	0.21	0.53	28	6
Size	0.16	0.50	18	78
Path-based Model Results				
Location	0.26	0.59	38	19
Reasoning	0.69	0.83	6	9
Sentiment	0.17	0.50	30	8
Size	0.16	0.53	16	5

Table 1: Results for Each KG Model

A separate model for each KG allows for controlling the types of templates to generate, as we try to reduce the skew in distribution presented in Figure 1. Triples are in the form of $P = (g(o_1), r_1, g(o_2))$. Moreover, using paths of $L > 1$ can allow us to generate questions about objects that are connected through intermediate nodes. We perform a grid search for each KG model to select the best model. Additionally, experimental details can be found in [1].

5.2. Experimental results

We use BLEU[3] and METEOR[4] scores to evaluate our question generation model. These metrics are commonly used for Natural Language Generation tasks. Table 1 presents the results of our triple-based ($L = 1$) and path-based ($L \leq 5$) models. Only the state gap path-based model achieves a higher METEOR score when compared to the triple-based model. Table 1 also contains the number of unique novel and existing (in the training data) templates that were generated. However, our state gap models do not perform well overall. We suspect that this might be because of the limited number of training templates (see [1]). Additionally, in [1], we show that our path-based direction gap model improves the METEOR score and BLEU score by 3% and 2% (respectively) when compared to the triple-based direction gap model. These results are understandable because questions regarding spatial relations can benefit from more relational information among objects.

Now, we present sample input and outputs from our test set using our following triple-based models: material, sentiment, reasoning, and size models. In the examples below, *I* stands for the input to the model, *GQT* marks the generated question templates, and *PQT* stands for the populated question template.

- I: *cap to the left of pants*
  GQT: "What is the OBJ near the OBJ made of ?"
  PQT: "What is the cap near the pants made of?"
- IN: *players to the right of man*
  GQT: "Which is younger, the OBJ or the OBJ ?"
  PQT: "Which is younger, the players or the man?"
- IN: *spectator to the right of cap*
  GQT: "Is the ATTRIBUTE OBJ to the right of the OBJ ?"
  PQT: "Is the happy spectator to the right of the cap?"
- IN: *bat to the right of shoe*
  GQT: "How big is the OBJ near the OBJ ?"
  PQT: "How big is the bat near the shoe?"

[3] https://www.nltk.org/_modules/nltk/translate/bleu_score.html

[4] https://www.nltk.org/_modules/nltk/translate/meteor_score.html

6. Conclusion

In this work, we take initial steps to understand the different types of reasoning skills needed for VQA tasks. We use a taxonomy of KGs to design a framework for identifying KGs for VQA questions. Using a question generation technique, we reduce the skew in the distribution of questions per KG category for the GQA dataset. In our future work, we aim to advance our question generation model and provide answers for generated questions. Additionally, we aim to generate more human-like question and to work towards resolving KGs. Our current work is an initial step towards understanding the cognitive skills need to advance AI agents.

Acknowledgements: We thank Dr. Wei-Lun Chao for suggestions on this work. The authors also acknowledge a seed grant from AFRL and OSU's Office of Research.

References

[1] Bajaj et al. Understanding knowledge gaps in visual question answering: Implications for gap identification and testing. *arXiv preprint arXiv:2004.03755*, 2020.

[2] Goyal et al. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *IEEE CVPR 2017*, pages 6904–6913.

[3] Gao et al. From two graphs to n questions: A vqa dataset for compositional reasoning on vision and commonsense. *arXiv preprint arXiv:1908.02962*, 2019.

[4] Hudson et al. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE CVPR 2019*, pages 6700–6709.

[5] Liu et al. Large-scale simple question generation by template-based seq2seq learning. In *NLPCC 2017*, pages 75–87.

[6] Marino et al. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *IEEE CVPR 2019*, pages 3195–3204, 2019.

[7] Wang et al. Fvqa: Fact-based visual question answering. *IEEE TPAMI*, 40(10):2413–2427, 2018.

[8] Yang et al. Visual curiosity: Learning to ask questions to learn visual recognition. In *CoRL 2018*, pages 63–80.

[9] Kushal Kafle and Christopher Kanan. Visual question answering: Datasets, algorithms, and future challenges. *CVIU*, 163:3–20, 2017.