Can Pre-trained Vision and Language Models Answer Visual Information-Seeking Questions? EMNLP 2023

Yang Chen[♠♡*] **Hexiang Hu**[♠] **Yi Luan**[♠] **Haitian Sun**[♠] **Soravit Changpinyo**[♣]
Alan Ritter[♡] **Ming-Wei Chang**[♠]
[♠]Google Deepmind [♣]Google Research [♡]Georgia Institute of Technology

Abstract

Pre-trained vision and language models (Chen et al., 2023b,a; Dai et al., 2023; Li et al., 2023b) have demonstrated state-of-the-art capabilities over existing tasks involving images and texts, including visual question answering. However, it remains unclear whether these models possess the capability to answer questions that are not only querying visual content but knowledge-intensive and information-seeking. In this study, we introduce INFOSEEK[1], a visual question answering dataset tailored for information-seeking questions that cannot be answered with only common sense knowledge. Using INFOSEEK, we analyze various pre-trained visual question answering models and gain insights into their characteristics. Our findings reveal that state-of-the-art pre-trained multi-modal models (e.g., PaLI-X, BLIP2, etc.) face challenges in answering visual information-seeking questions, but fine-tuning on the INFOSEEK dataset elicits models to use fine-grained knowledge that was learned during their pre-training. Furthermore, we show that accurate visual entity recognition can be used to improve performance on INFOSEEK by retrieving relevant documents, showing a significant space for improvement.

1 Introduction

The acquisition of knowledge occurs in the pre-training of large language models (Brown et al., 2020; Chowdhery et al., 2022), demonstrated as their emergent ability to answer information-seeking questions in the open-world, where the questioner does not have easy access to the information. While prior works have analyzed models' capabilities to answer *textual* information-seeking (or info-seeking) questions, much less is known for *visual* info-seeking questions. For example, after taking a picture of the specific church in Figure 1, a person might want to know the date of construction, or who decorated the interior of the church. Although the entity is presented in the image (the specific church), the relevant knowledge (e.g., the date) is not. Given recent advances on pre-trained visual and language models (Alayrac et al., 2022; Chen et al., 2023b; Li et al., 2023b), *do these models also understand how to answer visual information-seeking questions?*

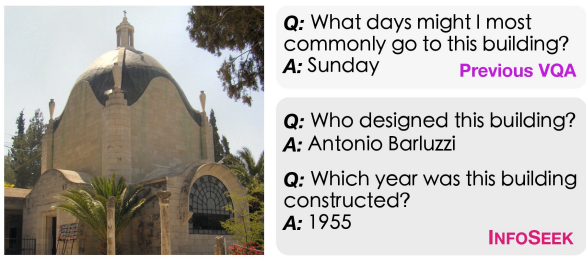

Figure 1: While 70.8% of OK-VQA questions can be answered by average adults without using a search engine, INFOSEEK poses challenges to query fine-grained information about the visual entity (e.g., `Dominus Flevit Church`), resulting in a sharp drop to 4.4% (§2).

To study this research question, a visual question answering (VQA) dataset focusing on info-seeking questions is inevitably required. However, not all VQA datasets meet this criterion. For example, by design, the majority of questions in datasets such as VQA v2 (Goyal et al., 2017) focus on visual attributes and object detection that does not require information beyond the image to answer. While models capable of answering these types of questions have the potential to aid visually impaired individuals (Gurari et al., 2018), there is a broader class of *info-seeking* questions that cannot be easily answered by sighted adults. Handling such questions (e.g., `When was this building constructed? 1955`) is critical as they come closer to the natural distribution of human questions.

In this paper, we present INFOSEEK, a natural

[*] Work done when interned at Google, yangc@gatech.edu

[1]Our dataset is available at https://open-vision-language.github.io/infoseek/.

VQA dataset that focuses on visual info-seeking questions. Different from previous VQA datasets, the testing subset of INFOSEEK is collected in multiple stages from human annotators to evaluate VQA where the question can not be answered from only the visual content (see a comparison of datasets in § 2). In addition to this manually curated test set, which enables realistic evaluation of info-seeking VQA, we also join annotations from a recent visual entity recognition dataset (Hu et al., 2023) with the Wikidata database (Vrandečić and Krötzsch, 2014), and employ human annotators to write templates to semi-automatically generate a large corpus of visual info-seeking QA pairs. Over 1 million {image, question, answer} triplets are generated to support fine-tuning multimodal models for info-seeking VQA. We split data to ensure memorizing knowledge during fine-tuning is useless — models either have to learn to use knowledge learned during pre-training or learn to retrieve knowledge from an external knowledge base.

Using INFOSEEK, we analyze the ability of state-of-the-art models to answer visual info-seeking questions. We found pre-trained vision-language models, such as models pre-trained end-to-end (e.g., PaLI-X by Chen et al.), and models pre-trained with frozen LLM (e.g., BLIP2 by Li et al.), both struggle to answer info-seeking questions in zero-shot, though BLIP2 outperforms PaLI-X by a margin. Surprisingly, after fine-tuning on our (large, semi-automatically curated) training set, PaLI-X yields a significant improvement and outperforms the fine-tuned BLIP2 models on queries that are unseen during fine-tuning. This suggests that while pre-trained PaLI-X has a significant amount of knowledge, it requires a small amount of fine-tuning data to fully awaken its capabilities. Furthermore, we show that INFOSEEK fine-tuned models can even generalize to questions and entity types completely unseen during fine-tuning (e.g., art & fashion).

When incorporating a visual entity recognition component, and conditioning models on the Wikipedia articles of the relevant entities, we show that models accessing such a knowledge base (With-KB) perform better overall than those that rely on knowledge learned during pre-training. However, end-to-end (No-KB) models were found better on certain classes of questions that require coarse-grained answers (*"Which continent is this building located on?"*), even on tail entities. Our experiment (§5.2) further suggests that improving visual entity recognition can drastically increase model's capability in answering visual info-seeking questions (from 18% to 45.6%), indicating a promising direction for future development.

Dataset	OK-VQA	ViQuAE	INFOSEEK
PaLM (Q-only)	23.8	**31.5**	5.6
Current SotA	66.1	22.1	18.2
Require Knowledge†	29.2%	95.2%	95.6%

† :% of questions that require knowledge to answer.
PaLM (Q-only): a question-only baseline using PaLM.

Table 1: Comparison of INFOSEEK and prior KI-VQA benchmarks. Performances reported in VQA score.

2 The Need for a New Visual Information-seeking Benchmark

While there have been plenty of knowledge-intensive VQA (KI-VQA) benchmarks, we show that none of these meet the criteria to effectively evaluate info-seeking VQA. Early efforts in this area, such as KBQA (Wang et al., 2015) and FVQA (Wang et al., 2017), were based on domain-specific knowledge graphs, while recent datasets like OK-VQA (Marino et al., 2019) and its variants such as S3VQA (Jain et al., 2021) and A-OKVQA (Schwenk et al., 2022) have improved upon this foundation by incorporating an open-domain approach and highlighting common-sense knowledge. Among the existing benchmarks, K-VQA (Sanket Shah and Talukdar, 2019) and ViQuAE (Lerner et al., 2022) are the most relevant, but they have severe limitations in their question generation process, as discussed below.

Information Seeking Intent. The evaluation of models' ability to answer info-seeking questions requires fine-grained knowledge, which a person is unlikely to know off the top of their head. However, we found that 70.8% of OK-VQA questions[2] can be answered without the need to use a search engine, indicating the dataset primarily focuses on knowledge that is commonly known to people. Most OK-VQA questions are regarding coarse-grained knowledge that many people already know: `What days might I most commonly go to this building? Sunday`. One only needs to know the building type (e.g., `Church`) rather than the specific building (e.g., `Dominus Flevit Church`). This

[2]Studied with human on 500 random OK-VQA questions (see Appendix C.1)