VQA-LOL: Visual Question Answering under the Lens of Logic

Tejas Gokhale* Pratyay Banerjee* Yezhou Yang Chitta Baral
Arizona State University, Tempe AZ
{tgokhale, pbanerj6, yz.yang, chitta}@asu.edu

Abstract

Logical connectives and their implications on the meaning of a natural language sentence are a fundamental aspect of understanding. In this paper, we investigate visual question answering (VQA) through the lens of logical transformation and posit that systems that seek to answer questions about images must be robust to these transformations of the question. If a VQA system is able to answer a question, it should also be able to answer the logical composition of questions. We analyze the performance of state-of-the-art models on the VQA task under these logical operations and show that they have difficulty in correctly answering such questions. We then construct an augmentation of the VQA dataset with questions containing logical operations and retrain the same models to establish a baseline. We further propose a novel methodology to train models to learn negation, conjunction, and disjunction and show improvement in learning logical composition and retaining performance on VQA. We suggest this work as a move towards embedding logical connectives in visual understanding, along with the benefits of robustness and generalizability. Our code and dataset is available online [1].

1. Introduction

Theories about logic in human understanding have a long history. In modern times, Piaget and Fodor [29] studied the development and representation of logical hypotheses in the human mind. Conjunction, disjunction, and negation were formalized into an "algebra of thought" by George Boole [5] as a way to improve, systemize, and mathematize Aristotle's Logic [10]. Horn considers negation to be a fundamental feature and a defining characteristic of human communication [16], following the traditions of Sankara [30], Spinoza [36], and Hegel [15]. Recent studies in [9] have suggested that infants can formulate intuitive and stable logical structures to interpret dynamic scenes and to entertain and rationally modify hypotheses about the scenes. As such we argue that understanding logical structures in questions, is a fundamental requirement for any question-answering system.

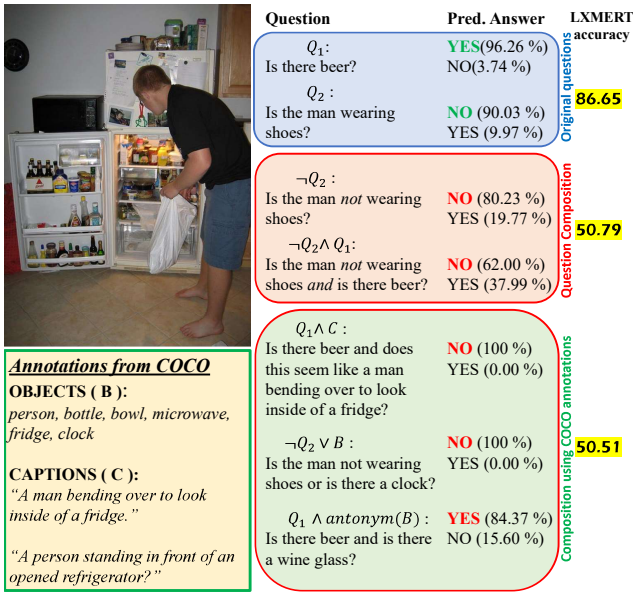

Figure 1. Illustration of logical composition of questions. The image and the questions Q_1 and Q_2 were taken from the VQA dataset. state-of-the-art models are able to answer Q_1 and Q_2 correctly, however predict the wrong answer when asked a logical composition such as negation ($\neg Q_1$), conjunction ($\neg Q_2 \wedge Q_1$), or disjunction ($\neg Q_2 \vee B$).

> If a question can be put at all, then it can be answered. [40]

In the above proposition, Wittgenstein linked the process of asking a question with the existence of an answer. While we do not comment on the existence of an answer, we suggest the following softer proposition -

> If a correct answer exists for a question Q, and if Q can be answered, then so should all the composite questions created from Q.

*Equal Contribution
[1] www.public.asu.edu/~tgokhale/vqa_lol.html