Commonsense Visual Question Answering for Autonomous Vehicle Images

Keegan Kimbrell
Department of Computer Science
University of Texas at Dallas
Richardson, Texas
kmk170830@utdallas.edu

2nd Gopal Gupta
Department of Computer Science
University of Texas at Dallas
Richardson, Texas
email address

3rd Kinjal Basu
Department of Computer Science
University of Texas at Dallas
Richardson, Texas
email address

***Abstract*—In this paper we propose a novel approach for visual question answering using commonsense knowledge and reasoning and apply it to images of roads taken for autonomous vehicles. Computer vision research is heavily reliant on machine learning techniques in a variety of areas, including visual question answering. We assert that logic programming can be applied to the output of a dense image captioning model to handle question answering. Commonsense reasoning allows us to create an explainable AI that comprehends and reasons over visual data in a manner similar to how a human would. This permits the answering of more complex and difficult questions that machine learning models fail to answer. Using logic programming comes at the added benefit of being readable and alterable, removing the need to train and test a new machine learning model when adjustments to the output must be made.**

We apply these concepts to a set of four images of streets to demonstrate how commonsense reasoning can be used to answer questions about autonomous vehicles. The program we made is able to determine if it safe for the driver to move forward, brake, change to the right lane, or change to the left lane given an image of the road in front of the driver. It does this by turning the visual data from the dense captioning model into logic predicates that allow it to determine if there are obstructions that prevent any of the given actions from occurring.

***Index Terms*—commonsense knowledge, commonsense reasoning, dense captioning**

I. Introduction

Natural Language question and answering is a pertinent area of interest in machine learning research, particularly when done in relation to image captioning. However, most machine learning models to encapsulate the complex relations between objects within a picture. Because of this, these image captioning models fail to answer particularly complicated questions about a picture that a person would have no problems with. Despite this, machine learning is the prevailing technique used in visual question answering. We believe that using logic programming to represent human knowledge and reason will allow us to more comprehensively answer natural language questions about that picture. This more closely emulates the manner in which people answer questions about a scene that they can see. We propose that we can apply a logic program to the results of an image captioning model to answer questions about an image.

In this paper, we will apply these concepts to create a program that answers a few questions about an image within the context of autonomous driving. The program receives an image of a road and returns safe actions for the driver to take. It does this by using turning the results of a *dense captioning* model into logic predicates. The program then feeds those predicates into a PROLOG logic program that represents the *commonsense reasoning* that a human would use to evaluate the picture. Finally, it returns safe driving actions based on the results of the logic program. The image captioning model used in this paper is called Dense Relational Captioning [1] which is a framework that generates captions the give information about the relations between objects identified in dense image regions. This paper also uses concepts from AUTO-DISCERN [2] and AQuA [3] to build the logic program used.

The remainder of the paper is as follows. Section 2 covers relating information and the work this paper is built upon. Section 3 outlines our methodology and lays out the full stack of programs used in the paper. Section 4 displays the results derived from running our program stack over four select images. Section 5 discusses the conclusions derived from our results and contemplates potential future work.

II. Background

There are numerous research projects dedicated to visual question answering, but many of these projects rely heavily on machine learning. In our approach, we instead use a dense captioning model and commonsense reasoning to perform this task. Here we present some of the major concepts and previous works that are used to achieve this.

A. Commonsense Reasoning

When humans view an image, they are able to make reasonable assumptions about the objects they are looking at extremely quickly. Several logic statements are made using the regions presented in the image to derive new information. In this manner, people use commonsense reasoning to generate the knowledge needed to answer questions about an image. To answer questions about an image like a human would, we need to simulate the commonsense reasoning and knowledge that a human has about a particular scene.

To cover the visual question answering part of our logic model, we borrow concepts from AQuA [3]. AQuA is a visual question answering framework that is capable of answering natural language questions about a picture. To relate objects and concepts that are similar, AQuA uses commonsense knowledge to define related attributes. AQuA performs this on simple 3D shapes, colors, and materials. However, this style of relating attribute information can be applied to pictures of real life scenes.

```
is_property(car, vehicle).
is_property(truck, vehicle).
is_property(suv, vehicle).
is_property(minivan, vehicle).
is_property(van, vehicle).
```

Defining the properties of objects and attributes as such allow us to create the commonsense knowledge needed to answer questions about a street or road. Our focus is on images of streets, particularly questions that can be used by an autonomous vehicle. We require commonsense reasoning about roads and street rules to improve our program's logic.

We use ideas presented in AUTO-DISCERN to create the commonsense reasoning needed to answer questions about street images [2]. The idea presented here is to take some information extracted from a computer vision model and apply that information into a set of logic rules that represent various factors on the road. The output will be an action that the driver should take given the information around them.

```
select_action(change_lane_right, T):-
\+obstruction_right(T).
select_action(change_lane_left, T):-
\+obstruction_left(T).
```

We implement a simple interpretation of the 'select action' logic presented in AUTO-DISCERN. This sets up the commonsense reasoning needed to determine legal actions on the road. As stated before, AUTO-DISCERN requires a computer vision model to provide it with the visual information it needs.

B. Dense Captioning

Computer vision is how programs represent the information within a picture as data and is largely handled with machine learning and deep learning models. In particular, image captioning is how these models can generate a natural language caption that represents the picture. To take this one step further, a new technique called dense captioning was proposed which generates natural language captions for multiple regions within an image. We require at least this much information to perform commonsense reasoning in a picture. This is because dense captioning gives us information of the objects within the picture, which allows us to reason over these objects.

The actual image captioning model used in this paper is Dense Relational Captioning [1]. This model improves upon the proposed dense captioning model by applying visual relationship detection. This gives us not only information about the objects being processed, but information about the relations between those objects. This is a very powerful addition to the knowledge base we are reasoning over, as it allows us to perform reasoning over the relationship of objects within a picture. For example, if we know that a fork is *on top* of a plate and that the plate is *on top* of a table, we can reason that the fork is *on top* or *above* the table. We can derive this information even if our Dense Relational Captioning model doesn't explicitly give it to us. For the autonomous vehicle concepts presented in this paper specifically, we will use this relational information to reason about the position of nearby objects on the road.

III. Methodology

Here, we provide the full stack of programs and methods used to perform this experiment.

A. Image Captioning

For each image provided to our program, we first run it through the Dense Relational Captioning model. This provides us with a JSON filled with visual information about the image, as seen in Figure 1. Included in the data created are the captions for the relations between dense regions and the coordinates of the bounding boxes that representing the captured regions. We manually annotate this JSON with another field that assigns to of the bounding boxes to each generated caption as demonstrated in Figure 2. This completes the visual information needed to move onto the logic section of the program.

B. Commonsense Reasoning

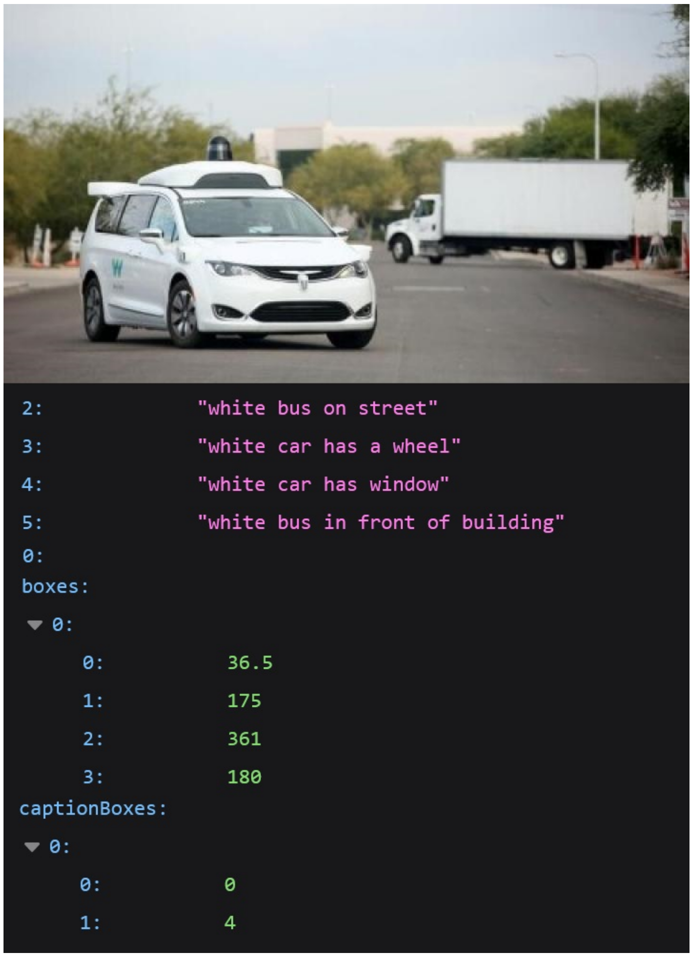
The visual information provided by the image captioning earlier is converted into logic predicate form. A program is run that uses spacy [4] to parse each natural language caption into two objects and a relation. It then attaches the manually added bounding boxes to the relation. Finally, it creates a list of separate predicates that represent each of the boxes and their coordinate information. This information is used as the input for a main PROLOG program that contains our commonsense knowledge and reasoning. Four queries are made for each image:

- `select_action(brake, T)`
- `select_action(change_lane_right, T)`
- `select_action(change_lane_left, T)`
- `select_action(go_forward, T)`

The program then returns a truth statement that represents whether or not each action is possible.

The main goal for each logic statement is to detect if there is an object or entity that is actively obstructing each action. Unlike most autonomous vehicles, we are operating with only an image and no additional information. To determine if it is safe to take an action, we look at the objects at different regions of the picture. We use commonsense knowledge and reasoning to determine if the detected objects are something that we can drive on (e.g. roads or streets) or are otherwise obstructing our path. This is done for all actions except for braking. For a brake to be safe, we simply determine if there

[1] Source of image: https://www.reuters.com/business/autos-transportation/us-labor-leader-calls-human-drivers-automated-vehicles-2021-05-17/.



```
2:            "white bus on street"
3:            "white car has a wheel"
4:            "white car has window"
5:            "white bus in front of building"
0:
boxes:
 ▼ 0:
     0:          36.5
     1:          175
     2:          361
     3:          180
captionBoxes:
 ▼ 0:
     0:          0
     1:          4
```

Fig. 1: A sample of the captions generated and one of the bounding boxes generated with Dense Relational Captioning model for the given image after manual annotation. The ”boxes” section contains four numbers represent the X-position, Y-position, width, and height of the bounding box. The ”captionBoxes” section states that the relation in caption zero is between region zero and region four. [1]

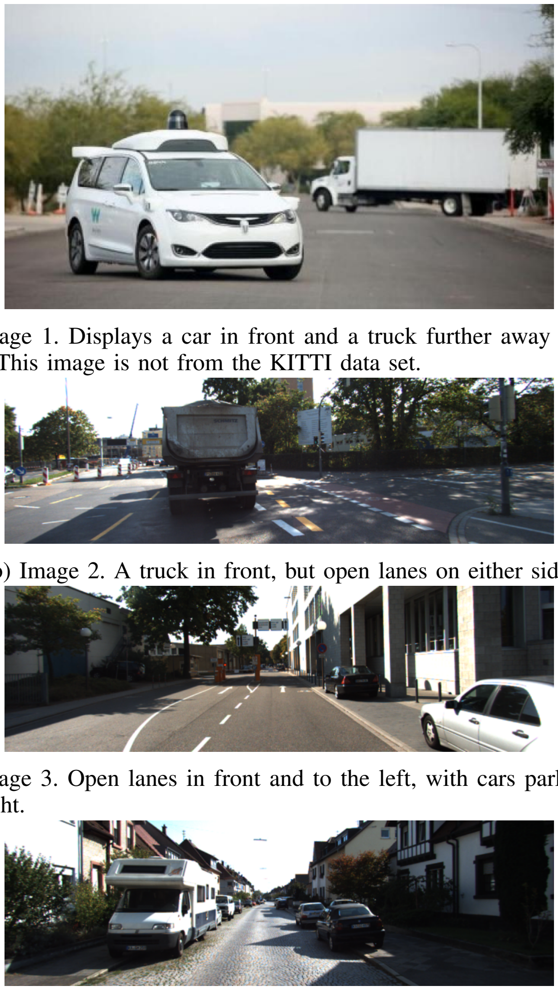
is something in our forward path and that it is expected of us to do so.

IV. Results

The program was ran over four images, one which was shown earlier and three more from the KITTI dataset [5]. As shown in Figure 2, these images show different road scenes in which multiple actions are safe and several aren’t.

A. Dense Captioning Results

We applied our dense captioning model to and created a visualization of the regions for these four images. This creates a group of bounding boxes as well as a list of captions for each image. As stated in our methodology, the output for each caption has the information for the related boxes manually attached to it.



(a) Image 1. Displays a car in front and a truck further away on the right. This image is not from the KITTI data set.

(b) Image 2. A truck in front, but open lanes on either side.

(c) Image 3. Open lanes in front and to the left, with cars parked on the right.

(d) Image 4. Open lane in front but vehicles parked on both sides.

Fig. 2: These four images depict different scenarios for street scenes. Images 2,3, and 4 were sourced from the KITTI data set.

As seen in Figure 3, the bounding boxes capture significant objects in the road scene. Smaller boxes represent more constrained objects such as vehicles, windows, or wheels. Larger boxes represent loosely confined objects such as streets, roads, or the background. A bounding box may represent a single object or multiple different objects depending on the captions generated by the dense captioning model. When a logic predicate considers a box, it considers all objects that the box represents as well.

B. Logic Query Results

Table 1 shows the results of turning the data collected from the dense captioning model into predicates and running our logic base. True means that an action is determined to be safe, while false means the opposite. This output demonstrates that our logic model generates correct answers for the images given here. It is important to note that due to the differing dimensions between the images from KITTI and the image that wasn’t,

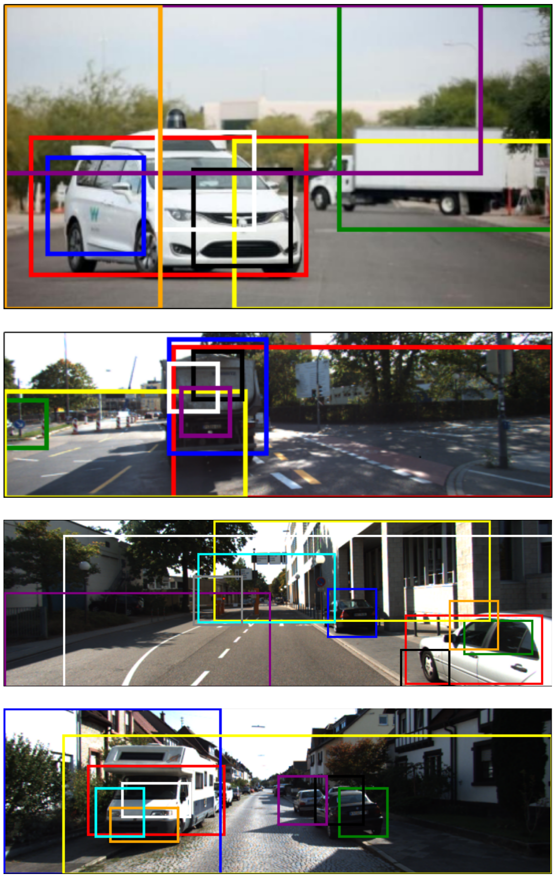


Fig. 3: The visual results of the dense captioning on the images. Each bounding box represents a region in the results.

there were changes to the parameters of the bounding boxes that evaluated the parameters. Despite this, we still generate accurate results and further proves the strength of having an alterable and readable artificial intelligence as opposed to using machine learning techniques.

TABLE I: Safe Actions per Image

Action	Image 1	Image 2	Image 3	Image 4
Go Forward	false	false	true	true
Change Lanes Left	false	true	true	false
Change Lanes Right	true	true	false	false
Brake	true	true	false	false

C. Logic Query Results

Following are the run times for the various sections of the experiment. The first two sections use machine learning while the third section uses only logic. The run time for the logic querying was consistently less than 1 ms.

TABLE II: Running Times in Seconds

Section	Image 1	Image 2	Image 3	Image 4
Dense Captioning	9.53	9.53	9.4	9.54
Predicate Generation	3.82	3.77	3.91	3.89
Querying	0	0	0	0

V. Conclusion

We proposed a new way of approaching visual question answering for images of real life scenery using commonsense knowledge and reasoning. The methodology was applied to the context of autonomous driving, and we showed how we can answer questions about safe actions on the road by reasoning over visual information like a human would. Our experiment demonstrates how the results of a computer vision model can be transformed into logic predicates. These predicates have been proven to be effective at answering questions about the image presented to it. Our approach is not only explainable and alterable, but it is also widely applicable. While we used this experiment to answer questions about autonomous driving, the techniques shown here can easily be applied to images of a different types. A more expansive knowledge base increases the variety of questions that can be asked, and more complex reasoning could allow for more complicated questions to be handled.

References

[1] D.-J. Kim, J. Choi, T.-H. Oh, and I. S. Kweon, "Dense relational captioning: Triple-stream networks for relationship-based captioning," 2019.

[2] S. Kothawade, V. Khandelwal, K. Basu, H. Wang, and G. Gupta, "Auto-discern: Autonomous driving using common sense reasoning," 2021.

[3] K. Basu, F. Shakerin, and G. Gupta, "Aqua: Asp-based visual question answering," pp. 57–72, 01 2020.

[4] M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing." To appear, 2017.

[5] J. Fritsch, T. Kuehnl, and A. Geiger, "A new performance measure and evaluation benchmark for road detection algorithms," in *International Conference on Intelligent Transportation Systems (ITSC)*, 2013.