READING ASSIGNMENT 1
16785-Integrated intelligence in robotics

[redacted]
Carnegie Mellon University

1. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning

As stated explicitly in the paper, a hand-crafted system with explicit knowledge of the CLEVR universe might work well, but will not generalize to real-world settings. The CLEVR universe consists of only three object shapes, two absolute sizes, two materials and eight colors. In order to improve this approach, images combinations covering an array of objects in the real world from trees to books to humans must be deployed. Also a form of contextual reasoning and semantic understanding should be deployed because it exists in the real world. Consider, a partially occluded apple in an image. Introduce an algorithm that recognizes the apple tree and by association realizes that the occluded object is indeed an apple.

Only 4 spatial relationships have been defined. More relationships can be defined in terms of distance from the object in question. This will reduce ambiguity in spatial relationships and help the model develop a cognitive approach to understanding spatial semantics.

Images in CLEVR are rendered using Blender. Blender is a biased rasterization engine, which means that it works by calculating which objects are visible to the camera. Using an engine such as Cycles is better. It simulates the the behavior of light by tracing it backwards, can also render caustics and is a better replica of the real world.

Decompose each question into sub-parts or modules and allocate a certain amount of long term memory to it. For example, 'how many brown cylinders exist?' Save the count of brown cylinders into the memory and retrieve it for the question, 'Are there more brown cylinders than red cylinders?' Since the count of brown cylinders already exists in the memory, computation is faster and comparisons have better results.

Currently the CLEVR dataset is being treated as a supervised learning problem by comparing the output with ground truth programs. However, it can be treated a semi supervised problem by using deep q learning or other reinforcement techniques to teach the algorithm the best program for each family of questions. This may also help reduce biases learning during training and testing.

2. Unbiased Look at Dataset Bias

This paper proposes cross dataset generalization as a means of negating bias. The method proposed is outdated as it was set in 2011. Presumably deep neural networks will perform much better across datasets as they will work towards minimizing the test loss function.

Category or label bias has been introduced. One way of reducing it is by assigning all the synonyms of a particular label to a certain category during training i.e. a category can hold more than one title. For example, images of rabbits will be assigned to the same category holding the labels, rabbit and bunny.

Capture bias can be reduced by warping the image in question. For example, a single dataset may have learn to recognize a car from its front view alone because there aren't sufficient images of the rear during training. Perform linear transformations such as inversion on these training images is most effective. Also, swap in objects in certain images with objects from others.

Try to render images synthetically. Jitter the light and camera viewpoint randomly in each image. This automatically eliminates selection bias and capture bias found in images online.

Eliminate the use of outdated datasets such as Caltech-101 and MSRC. Focus on building algorithm that are successful for the bigger picture and that perform better across varied datasets. Avoid fine tuning an algorithm to overfit a single dataset. Extend the experimental setup described in the paper to applications beyond the visual world. For example, speech recognition in devices like Alexa. They tend to understand the dialect of the western world a lot faster than the eastern world. Reducing the accent based bias in this scenario would be helpful.