Visual Reasoning with Natural Language

Stephanie Zhou∗, Alane Suhr∗, and Yoav Artzi
Dept. of Computer Science and Cornell Tech
Cornell University
New York, NY 10044
sz244@cornell.edu suhr@cs.cornell.edu yoav@cs.cornell.edu

Introduction

Natural language provides a widely accessible and expressive interface for robotic agents. To understand language in complex environments, agents must reason about the full range of language inputs and their correspondence to the world. For example, consider the scenario and instruction in Figure 1. To execute the instruction, the robot must identify the top shelf, recognize the two stacks as sets of items, compare items, and reason about the content and size of the sets. Such reasoning over language and vision is an open problem that is receiving increasing attention (Antol et al. 2015; Chen et al. 2015; Johnson et al. 2016). While existing data sets focus on visual diversity, they do not display the full range of natural language expressions, such as counting, set reasoning, and comparisons.

We propose a simple natural language visual reasoning task, where the goal is to predict if a descriptive statement paired with an image is true for the image. This abstract describes our existing synthetic images corpus (Suhr et al. 2017) and current work on collecting real vision data.

Related Work

Several tasks focus on language understanding in visual contexts, including caption generation (Chen et al. 2015; Young et al. 2014; Plummer et al. 2015), visual question answering (Antol et al. 2015), referring expression resolution (Matuszek et al. 2012; Krishnamurthy and Kollar 2013) and generation (Mitchell, van Deemter, and Reiter 2010; FitzGerald, Artzi, and Zettlemoyer 2013), and mapping of instructions to actions (MacMahon, Stankiewics, and Kuipers 2006; Chen and Mooney 2011; Artzi and Zettlemoyer 2013; Bisk, Yuret, and Marcu 2016; Misra, Langford, and Artzi 2017). We focus on visual reasoning with emphasis on linguistic diversity. The most related resource to ours is CLEVR (Johnson et al. 2016), where questions are paired with synthetic images. However, in contrast to our work, both language and images are synthetic.

∗Contributed equally.

Copyright © 2017, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
Krippendorff’s for each pair, and observe high inter-annotation agreement and development sets we collect five validation judgements unreleased and used for the task leaderboard. For testing we create four sets: one test set is public, and the second is 80.7% for training, 6.4% for development, and the rest for total of 92,244 sentence-image pairs. We create four sets: 

Data Collection We define a two-stage process: sentence writing and validation. Figure 3 illustrates the sentence writing stage. This task requires workers to identify similarities and differences between images, and requires careful reasoning, which is reflected in the collected language. We show workers four generated images, each made of three boxes containing shapes. The first two images are generated independently. The third and fourth are generated from the first and second by shuffling objects. This discourages trivial sentences, such as there is a blue triangle. We ask for a sentence that is true for the first two images, and false for the others. We instruct workers that sentences may not refer to the order of boxes. This enables permuting the boxes while retaining the statement truth value. We pair each image with the written sentence to create four pairs. In the validation stage, we ask for a label for each pair. While the truth-value can be inferred from the sentence-writing stage, validation increases data quality. Finally, we generate six image-sentence pairs by permuting the three boxes in each image.

Data Statistics We collect 3,962 unique sentences for a total of 92,244 sentence-image pairs. We create four sets: 80.7% for training, 6.4% for development, and the rest for two test sets. One test set is public, and the second is unreleased and used for the task leaderboard. For testing and development sets we collect five validation judgements for each pair, and observe high inter-annotation agreement (Krippendorff’s $\alpha = 0.31$ and Fleiss’ $\kappa = 0.808$).

Analysis We analyze our corpus and existing corpora for linguistic complexity. We classify for a broad set of linguistic phenomena, including quantification, cardinal reasoning, syntactic ambiguity, and semantic and pragmatic features (e.g., coreference and spatial relations). The details of the analysis are in Suhr et al. (2017). Our analysis shows our data is significantly more linguistically diverse than VQA. For example, 66% of our sentences refer to exact counts, whereas this occurs in only 12% of sentences in VQA.

Baselines We evaluate the difficulty of the task with multiple baselines. We construct two models that use only one of the input modalities to measure biases. Both perform similarly to the majority-class baselines. The best-performing model is neural module networks (Andreas et al. 2016), which achieves 62% on the unreleased test set. The original publication includes a break down of sampled errors per analyzed linguistic phenomena.

Real Vision Data

We are currently collecting an NLVR real vision data set. Our goal is to collect statements displaying a variety of linguistic phenomena, such as counting, spatial relations, and comparisons. In contrast to our use of synthetic images, we aim for realistic visual input, including a broad set of object types and scenes. Figure 4 shows initial examples. To correctly reason about the top statement, the system must maximize a spatial property and identify the number of images in which it holds. To understand the second statement, the agent has to consider several unique objects and compare a certain property they all demonstrate—the direction they face.

Conclusion

We describe a task for language and vision reasoning, and a newly released vision and language data set, the Cornell Natural Language Visual Reasoning (NLVR) corpus. While the task is straightforward to evaluate, it requires complex reasoning. Performance of existing methods demonstrates the challenge the data presents. We also discuss ongoing work on collecting similar data that includes both linguistically diverse text and real vision challenges.

Write one sentence. This sentence must meet all of the following requirements:

- It describes A.
- It describes B.
- It does not describe C.
- It does not describe D.
- It does not mention the images explicitly (e.g. “In image A, ...”).
- It does not mention the order of the light grey squares (e.g. “In the rightmost square(...)”)

There is no one correct sentence for this image. There may be multiple sentences which satisfy the above requirements. If you can think of more than one sentence, submit only one.

Figure 3: Sentence writing prompt. The top sentence in Figure 2 was generated from this prompt.

Figure 4: Crowdsourced sentences and images from our ongoing work. The truth value of the top sentence is true, while the bottom is false.

A red vest is furthest to the left in at least one paired image.

The gazelle in both pictures are running the same direction.
References


