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.

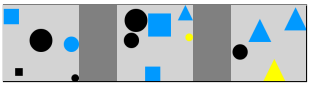
Related Work

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.

Fold and place the sweatshirt on the top shelf, and make sure the stacks are the same color and evenly distributed.

Figure 1: An example instruction that may be given to a household robot.



There is a box with 2 triangles of same color nearly touching each other.

There are two towers with the same height but their base is not the same in color.

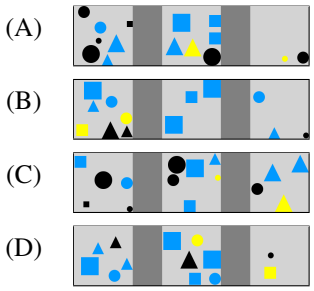
Figure 2: Example for natural language visual reasoning. The top sentence is false, while the bottom is true.

Task

Given an image and a natural language statement, the task is to predict whether the statement is true in regard to the image. Figure 2 shows two examples with generated images. The statement in the top example is true in regard to the given image, while the lower example is false. We evaluate system performance using accuracy. This provides a straightforward evaluation metric, in contrast to other related tasks, which use partial credit metrics, such as BLEU.

Synthetic Image Data

In Suhr et al. (2017), we present the Cornell Natural Language Visual Reasoning (NLVR) corpus. The corpus includes statements paired with synthetic images. Using synthetic images enables control of the visual content and reasoning required to distinguish between images. We briefly



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.

review the data, collection process, and considerations, and refer to the original publication for the details.

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' $\kappa = 0.808$).

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

The gazelle in both pictures are running the same direction

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

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.

References

Andreas, J.; Rohrbach, M.; Darrell, T.; and Klein, D. 2016. Neural module networks. In *Conference on Computer Vision and Pattern Recognition*.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: Visual question answering. In *International Journal of Computer Vision*.

Artzi, Y., and Zettlemoyer, L. 2013. Weakly supervised learning of semantic parsers for mapping instructions to actions. *Transactions of the Association of Computational Linguistics* 1:49–62.

Bisk, Y.; Yuret, D.; and Marcu, D. 2016. Natural language communication with robots. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Chen, D. L., and Mooney, R. J. 2011. Learning to interpret natural language navigation instructions from observations. In *Proceedings of the National Conference on Artificial Intelligence*.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft COCO captions: Data collection and evaluation server. *CoRR* abs/1504.00325.

FitzGerald, N.; Artzi, Y.; and Zettlemoyer, L. 2013. Learning distributions over logical forms for referring expression generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Johnson, J.; Hariharan, B.; van der Maaten, L.; Fei-Fei, L.; Zitnick, C. L.; and Girshick, R. B. 2016. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. *CoRR* abs/1612.06890.

Krishnamurthy, J., and Kollar, T. 2013. Jointly learning to parse and perceive: Connecting natural language to the physical world. *Transactions of the Association for Computational Linguistics* 1.

MacMahon, M.; Stankiewics, B.; and Kuipers, B. 2006. Walk the talk: Connecting language, knowledge, action in route instructions. In *Proceedings of the National Conference on Artificial Intelligence*.

Matuszek, C.; FitzGerald, N.; Zettlemoyer, L. S.; Bo, L.; and Fox, D. 2012. A joint model of language and perception for grounded attribute learning. In *Proceedings of the International Conference on Machine Learning*.

Misra, D.; Langford, J.; and Artzi, Y. 2017. Mapping instructions and visual observations to actions with reinforcement learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Mitchell, M.; van Deemter, K.; and Reiter, E. 2010. Natural reference to objects in a visual domain. In *International Conference on Natural Language Generation*.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *The IEEE International Conference on Computer Vision*.

Suhr, A.; Lewis, M.; Yeh, J.; and Artzi, Y. 2017. A corpus of natural language for visual reasoning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations. *Transactions of the Association of Computational Linguistics*.