A Corpus for Reasoning About Natural Language Grounded in Photographs

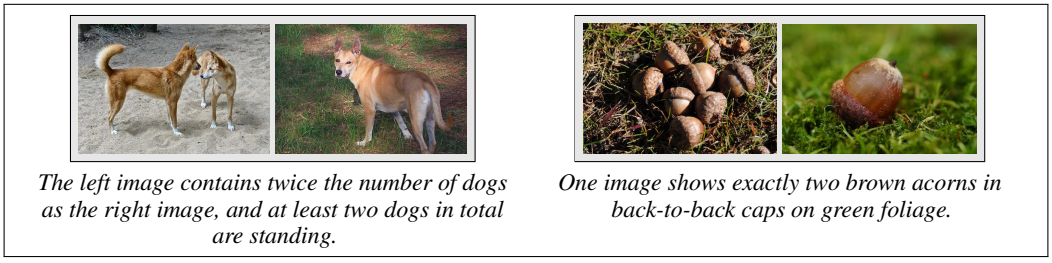
Alane Suhr[‡,*], **Stephanie Zhou**[†,*], **Iris Zhang**[‡], **Huajun Bai**[‡], and **Yoav Artzi**[‡]

[‡]Department of Computer Science and Cornell Tech, Cornell University
New York, NY, 10044
{suhr, yoav}@cs.cornell.edu {wz337, hb364}@cornell.edu

[†]Department of Computer Science, University of Maryland
College Park, MD 20742
stezhou@cs.umd.edu

Abstract

We introduce a new dataset for the task of deciding whether a caption is true about an image. The data contains 107,296 examples of 29,680 unique English sentences paired with natural photographs. We present an approach for finding visually complex images and crowdsourcing linguistically diverse captions. Qualitative analysis shows the data requires compositional reasoning about quantities, comparisons, and spatial relations. Human performance and evaluation of state-of-the-art visual reasoning methods show the data is an open challenge for current methods.



The left image contains twice the number of dogs as the right image, and at least two dogs in total are standing.

One image shows exactly two brown acorns in back-to-back caps on green foliage.

Figure 1: Two examples from NLVR[2]. Each caption is paired with two images. The task is to predict if the caption is True or False. The truth value of the left sentence is True, while the right is False.[1]

1 Introduction

Jointly reasoning about linguistic and visual inputs is receiving increasing research attention. However, the challenges presented by available resources [e.g., Zitnick and Parikh, 2013, Antol et al., 2015, Chen et al., 2016] are far from reflecting the full complexity of the problem. The language is often relatively simple, requiring reasoning that is not fundamentally different than traditional vision problems, often consisting of only identifying simple properties of an object or a small set of spatial relations between two objects. This motivated the design of more complex visual reasoning tasks, including NLVR [Suhr et al., 2017] and CLEVR [Johnson et al., 2017a,b]. However, both tasks use synthetic images, and the most widely studied version of CLEVR uses synthetic language as well. As a result, these resources only partially reflect the challenges of language and vision.

We study the challenges of jointly reasoning about language and vision by identifying visual properties that enable compositional reasoning and the type of language they bring about. We construct Natural

[*]Contributed equally.
[†]Work done as an undergraduate at Cornell University.
[1]The Supplementary Material contains license information on the photographs in this paper.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

Language Visual Reasoning for Real (NLVR^2), a new dataset focused on web photos for the task of determining if a statement is true with regard to an image. Our analysis shows that joint reasoning about complex visual input and diverse language requires compositional reasoning, including about sets, properties, counts, comparisons, and spatial relations. Figure 1 shows examples from NLVR^2.

Scalable curation of language and vision data that requires complex reasoning requires addressing two challenges. First, we must identify images that have sufficient visual complexity to allow for the type of reasoning desired. For example, a photo of a single beetle with a uniform background supports limited reasoning beyond the existence of the beetle and its properties. Second, we need a scalable process to collect a large set of captions that are linguistically diverse. We query a search engine with queries designed to yield visually complex photographs. Rather than presenting workers with a single image to annotate, we elicit interesting captions by asking workers to compare and contrast four pairs of images at the same time, a process inspired by Suhr et al. [2017]. The caption must be true for two pairs, and false for the others. Finally, we generate an example for each pair for the task of determining if a caption is true for the pair.

NLVR^2 contains $107{,}296$ examples, each of a caption and an image pair. The data includes $29{,}680$ unique sentences and $127{,}506$ images. Qualitative linguistic analysis shows that our data retains the broad representation of linguistic phenomena seen in NLVR, while displaying a far more interesting vision challenge. We evaluate the challenge that NLVR^2 presents using a set of baselines and state-of-the-art visual reasoning methods. The relatively low performance shows that NLVR^2 presents a significant challenge even for methods that perform well on existing visual reasoning tasks, demonstrating how using natural inputs for both modalities better exposes the challenges of the task.

2 Related Work and Datasets

Natural language understanding in the context of images has been studied within various tasks, including visual question answering [Zitnick and Parikh, 2013, Antol et al., 2015], caption generation [Chen et al., 2016], referring expression resolution [Mitchell et al., 2010, Matuszek et al., 2012, FitzGerald et al., 2013], and instruction following [MacMahon et al., 2006, Chen and Mooney, 2011, Bisk et al., 2016, Misra et al., 2018, Blukis et al., 2018]. Several recent datasets focus on compositional reasoning about images and language, mostly using synthetic data for both language and vision [Andreas et al., 2016b, Johnson et al., 2017a,b, Kuhnle and Copestake, 2017, Kahou et al., 2018, Yang et al., 2018]. Two exceptions are CLEVR-Humans [Johnson et al., 2017b] and NLVR [Suhr et al., 2017], which use crowdsourced language data. Several methods have been proposed for compositional visual reasoning [Andreas et al., 2016a,b, Johnson et al., 2017b, Perez et al., 2018, Hu et al., 2017, Mascharka et al., 2018, Hu et al., 2018, Suarez et al., 2018, Santoro et al., 2017, Zhu et al., 2017, Hudson and Manning, 2018, Tan and Bansal, 2018, Malinowski et al., 2018]. In contrast to synthetic images or text, we focus both on human-written language and web photographs. Our approach is inspired by the collection of NLVR, where workers were shown a set of similar images and asked to write a sentence true for some images, but false for the others. This requires workers to compare and constrast sets of photographs. We adapt this method to web photos, including introducing a process to identify images that support complex reasoning.

3 Collecting NLVR^2

Each example in NLVR^2 includes a pair of images and a natural language sentence. The task is to determine whether the sentence is True or False about the pair of images. This binary prediction task allows a straightforward evaluation of accuracy. We design a process to identify images that enable complex reasoning, collect grounded natural language descriptions, and label them as True or False. Figure 2 illustrates the collection procedure, which includes a sequence of four crowdsourcing tasks.

We use 124 synsets from the ILSVRC2014 ImageNet challenge [Russakovsky et al., 2015] to generate search queries and retrieve sets of images with similar content. The synset correspondence allows use of models pre-trained on ImageNet, such as ResNet [He et al., 2016] or Inception [Szegedy et al., 2016]. We begin by collecting sets of eight images for each synset, through a three-step process: generating search queries and downloading sets of similar images (Figure 2a), pruning low-quality images (Figure 2b), and constructing sets of eight images (Figure 2c). We display randomly paired images from each set of eight to a worker, and ask them to select two pairs, and write a sentence that is true about the selected pairs but false about the others (Figure 2d). This requires workers to find similarities and differences between sets of images, which encourages more compositional

(a) **Find sets of images:** For each synset, we generate a search queries designed to retrieve visually complex images. We combine synset names with numerical expressions, hypernyms in WordNet [Miller, 1993, Deng et al., 2014, Russakovsky et al., 2015], words close in the embedding space of word2vec [Mikolov et al., 2013], and activities. In this example, the query `two acorns` is generated for the synset `acorn` and is issued to the search engine. The leftmost image appears in the list of results. The similar images tool is used to find a set of images, shown on the right, that are similar to this image.

(b) **Image pruning:** We remove images identified as low-quality by crowdworkers, including images that do not contain the synset, images containing inappropriate content, or non-realistic artwork or collages. In this example, one image is removed because it does not show an `acorn`.

(c) **Set construction:** Crowdworkers mark the remaining images as interesting or non-interesting. We define interesting images as showing more than one instance of the synset, showing the synset interacting with other objects, showing the synset performing an activity, or displaying a set of diverse objects or features. In this example, three images are marked as non-interesting (top row) because they contain only a single instance of the synset. Any sets with fewer than three interesting images are discarded. The images are re-ordered (bottom row) so that interesting images appear before non-interesting images, and the top eight images are used to form the set. In this example, the set is formed using the leftmost eight images.

(d) **Sentence writing:** The images in the set are randomly paired and shown to the worker. The worker selects two pairs, and writes a sentence that is true for the two selected pairs but false for the other two pairs.

(e) **Validation:** The set is split into four examples, where each example consists of the sentence and a pair of images. Each example is shown to a worker, who labels it as True or False.

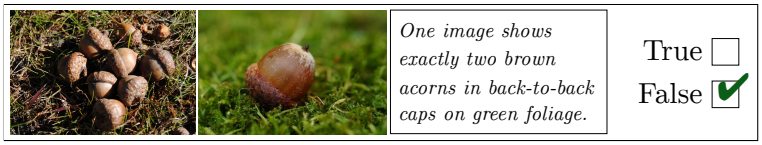


Figure 2: Illustration of the data collection process, showing construction of a single example.

sentences [Suhr et al., 2017]. Our guidelines encourage more compositionally challenging sentences. Each set of eight images is used for two sentence-writing tasks. Finally, we split each sentence-writing task into four examples, where the sentence is paired with each pair of images. We show each example independently to a worker, who labels the example as True or False (Figure 2e).

We collect additional validation judgments for randomly selected 20% of the examples, ensuring that examples from the same initial set of eight images do not appear across the split. For each of these examples, we collect an additional four validation judgments, and remove examples where two or more workers did not agree with the original label. We split this 20% into three equal-sized splits to form the development and two test sets. The rest is used as training data. One of the test sets is unreleased and will be used for the task leaderboard only. In total, NLVR^2 consists of 107,296 examples, covering 29,680 unique sentences and 107,296 unique images. The total cost of data

	VQA	NLVR	NLVR^2	Example from NLVR^2
Cardinality (hard)	11.5	66	28	***Six*** *rolls of paper towels are enclosed in a plastic package with the brand name on it.*
Cardinality (soft)	1	16	22.5	*In* ***at least one image*** *there is a bottle of wine and a glass with red wine in it.*
Existential	11.5	88	21.5	***There are*** *at most 3 water buffalos in the image pair.*
Universal	1	7.5	18	***Each image*** *contains exactly one eagle with outspread wings flying in a clear blue sky.*
Coordination	5	17	28	*Each image contains only one wolf,* ***and*** *all images include snowy backdrops.*
Coreference	6.5	3	14.5	*there are four or more animals very close to* ***each other*** *on the grass in the image to the left.*
Spatial Relations	42.5	66	56.5	*At least one panda is sitting* ***near*** *a fallen branch* ***on*** *the ground.*
Comparative	1	3	9	*There are* ***more*** *birds in the image on the right* ***than*** *in the image on the left.*
Presupposition	80	19.5	16	*A cookie sits in* ***the dessert*** *in the image on the left.*
Negation	1	9.5	13.5	*The front paws of the dog in the image on the left are* ***not touching*** *the ground.*

Table 1: Linguistic analysis of sentences from NLVR^2, VQA, and NLVR. We analyze 200 sentences from each dataset for presence of the semantic phenomena described in Suhr et al. [2017].

collection was $19,132.99. The inter-annotator agreement, measured over the development and test sets, is near perfect (Krippendorff's $\alpha = 0.91$ and Fleiss' $\kappa = 0.89$).

4 Data Analysis

We study the reasoning challenge of the data by analyzing sentences for presence of the linguistic phenomena identified in Suhr et al. [2017]. Table 1 summarizes the results of this analysis, comparing NLVR^2 with VQA Antol et al. [2015] and NLVR Suhr et al. [2017]. NLVR^2 presents a significant visual reasoning challenge, requiring a language understanding system to be able to count, compare objects, reason about sets, and understand compositional semantic phenomena such as negation and coordination. For many phenomena, such as coordination, comparisons, and negation, NLVR^2 contains more examples than NLVR or VQA. This demonstrates how our use of real images and the data collection procedure elicits more diverse natural language.

5 Experiments and Results

We evaluate the difficulty of NLVR^2 using baselines and state-of-the-art visual reasoning methods. We measure accuracy as the proportion of examples for which the model predicts the correct truth value. We report accuracies on the unreleased test set. Human accuracy is 96.1%. The data is balanced between True and False labels; majority class accuracy predicting True results in 51.4% accuracy. We construct two baselines that use only one of the input modalities to measure whether the task can be solved without observing both the text and image. These approaches result in accuracy near majority class, showing that both modalities must be used to solve the task. The highest accuracy, 53.5%, is achieved by training a maximum entropy classifier with features computed by combining results of object detection over the images [He et al., 2017, Girshick et al., 2018] and numerical expressions in the sentence. We evaluate using several methods that achieved state-of-the-art results on the CLEVR dataset: N2NMN [Hu et al., 2017], FiLM [Perez et al., 2018], and MAC-Network [Hudson and Manning, 2018]. We adapt these methods to process pairs of images by computing a single set of image features for the two images concatenated horizontally. N2NMN and MAC-Network perform near majority class accuracy, while FiLM achieves 53.0% accuracy.

6 Discussion

We introduce the NLVR^2 corpus for studying joint reasoning about photographs and natural language captions. Our analysis shows that the language contains a range of linguistic phenomena including numerical expressions, quantifiers, coreference, and negation. Our experimental results and our analysis illustrate the challenge that NLVR^2 introduces to methods for visual reasoning. We release training, development, and public test sets. Procedures for evaluating on the unreleased test set and a leaderboard will be available at `https://github.com/clic-lab/nlvr`.

References

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Learning to compose neural networks for question answering. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1545–1554, 2016a. doi: 10.18653/v1/N16-1181. URL http://www.aclweb.org/anthology/N16-1181.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 39–48, 2016b. doi: 10.1109/CVPR.2016.12.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *IEEE International Conference on Computer Vision*, pages 2425–2433, 2015. doi: 10.1109/ICCV.2015.279.

Yonatan Bisk, Daniel Marcu, and William Wong. Towards a dataset for human computer communication via grounded language acquisition. In *Proceedings of the AAAI Workshop on Symbiotic Cognitive Systems*, 2016.

Valts Blukis, Dipendra Misra, Ross A. Knepper, and Yoav Artzi. Mapping navigation instructions to continuous control actions with position visitation prediction. In *Proceedings of the Conference on Robot Learning*, 2018.

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

Wenhu Chen, Aurélien Lucchi, and Thomas Hofmann. Bootstrap, review, decode: Using out-of-domain textual data to improve image captioning. *CoRR*, abs/1611.05321, 2016. URL http://arxiv.org/abs/1611.05321.

Jia Deng, Olga Russakovsky, Jonathan Krause, Michael S. Bernstein, Alex Berg, and Li Fei-Fei. Scalable multi-label annotation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3099–3102, 2014. doi: 10.1145/2556288.2557011. URL http://doi.acm.org/10.1145/2556288.2557011.

Nicholas FitzGerald, Yoav Artzi, and Luke Zettlemoyer. Learning distributions over logical forms for referring expression generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1914–1925, 2013. URL http://www.aclweb.org/anthology/D13-1197.

Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. Detectron. https://github.com/facebookresearch/detectron, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask r-cnn. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. *IEEE International Conference on Computer Vision*, pages 804–813, 2017.

Ronghang Hu, Jacob Andreas, Trevor Darrell, and Kate Saenko. Explainable neural computation via stack neural module networks. *European Conference on Computer Vision*, 2018.

Drew A. Hudson and Christopher D. Manning. Compositional attention networks for machine reasoning. In *Proceedings of the International Conference on Learning Representations*, 2018.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1988–1997, 2017a.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. Inferring and executing programs for visual reasoning. *IEEE International Conference on Computer Vision*, pages 3008–3017, 2017b.

Samira Ebrahimi Kahou, Adam Atkinson, Vincent Michalski, Ákos Kádár, Adam Trischler, and Yoshua Bengio. FigureQA: An annotated figure dataset for visual reasoning. In *Proceedings of the International Conference on Learning Representations*, 2018.

Alexander Kuhnle and Ann A. Copestake. ShapeWorld - a new test methodology for multimodal language understanding. *CoRR*, abs/1704.04517, 2017.

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

Mateusz Malinowski, Carl Doersch, Adam Santoro, and Peter Battaglia. Learning visual question answering by bootstrapping hard attention. *CoRR*, abs/1808.00300, 2018.

David Mascharka, Philip Tran, Ryan Soklaski, and Arjun Majumdar. Transparency by design: Closing the gap between performance and interpretability in visual reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Cynthia Matuszek, Nicholas FitzGerald, Luke S. Zettlemoyer, Liefeng Bo, and Dieter Fox. A joint model of language and perception for grounded attribute learning. In *Proceedings of the International Conference on Machine Learning*, 2012.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR, abs/1301.3781*, 2013.

George A. Miller. WordNet: A lexical database for English. In *Proceedings of the Workshop on Human Language Technology*, pages 409–409, 1993. ISBN 1-55860-324-7. doi: 10.3115/1075671.1075788. URL https://doi.org/10.3115/1075671.1075788.

Dipendra Misra, Andrew Bennett, Valts Blukis, Eyvind Niklasson, Max Shatkhin, and Yoav Artzi. Mapping instructions to actions in 3D environments with visual goal prediction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018.

Margaret Mitchell, Kees van Deemter, and Ehud Reiter. Natural reference to objects in a visual domain. In *Proceedings of the International Natural Language Generation Conference*, 2010. URL http://www.aclweb.org/anthology/W10-4210.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. FiLM: Visual reasoning with a general conditioning layer. In *AAAI Conference on Artificial Intelligence*, 2018.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3): 211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Adam Santoro, David Raposo, David G.T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems*, pages 4967–4976, 2017.

Joseph Suarez, Justin Johnson, and Fei-Fei Li. DDRprog: A CLEVR differentiable dynamic reasoning programmer. *CoRR*, abs/1803.11361, 2018. URL http://arxiv.org/abs/1803.11361.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 217–223, 2017. doi: 10.18653/v1/P17-2034. URL http://www.aclweb.org/anthology/P17-2034.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex A. Alemi. Inception-v4, Inception-ResNet and the impact of residual connections on learning. In *Proceedings of the International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1602.07261.

Hao Tan and Mohit Bansal. Object ordering with bidirectional matchings for visual reasoning. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 444–451, 2018. URL http://aclweb.org/anthology/N18-2071.

Robert Guangyu Yang, Igor Ganichev, Xiao Jing Wang, Jonathon Shlens, and David Sussillo. A dataset and architecture for visual reasoning with a working memory. In *European Conference on Computer Vision*, 2018.

Chen Zhu, Yanpeng Zhao, Shuaiyi Huang, Kewei Tu, and Yi Ma. Structured attentions for visual question answering. *IEEE International Conference on Computer Vision*, pages 1300–1309, 2017.

C. Lawrence Zitnick and Devi Parikh. Bringing semantics into focus using visual abstraction. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3009–3016, 2013. doi: 10.1109/CVPR.2013.387.

A License Information

Tables 2 and 3 contain license and attribution information on the images included in this paper.

Image	Attribution and License
	MemoryCatcher (CC0)
	Calabash13 (CC BY-SA 3.0)
	Albert Bridge (CC BY-SA 2.0)
	Randwick (CC BY-SA 3.0)

Table 2: License information for the images in Figure 1.

Image	Attribution and License
	Hagerty Ryan, USFWS (CC0)
	Charles Rondeau (CC0)
	Peter Griffin (CC0)
	Petr Kratochvil (CC0)
	George Hodan (CC0)
	Charles Rondeau (CC0)
	Andale (CC0)
	Maksym Pyrizhok (PDP)
	Sheila Brown (CC0)
	ulleo (CC0)

Table 3: License information for the images in Figure 2.