A Corpus for Reasoning About Natural Language
Grounded in Photographs

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

[‡]Cornell University Department of Computer Science and Cornell Tech
New York, NY 10044
{suhr, yoav}@cs.cornell.edu   {az346, wz337, hb364}@cornell.edu

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

Abstract

We introduce a new dataset for joint reasoning about natural language and images, with a focus on semantic diversity, compositionality, and visual reasoning challenges. The data contains 107,292 examples of English sentences paired with web photographs. The task is to determine whether a natural language caption is true about a pair of photographs. We crowdsource the data using sets of visually rich images and a compare-and-contrast task to elicit linguistically diverse language. Qualitative analysis shows the data requires compositional joint reasoning, including about quantities, comparisons, and relations. Evaluation using state-of-the-art visual reasoning methods shows the data presents a strong challenge.

1 Introduction

Visual reasoning with natural language is a promising avenue to study compositional semantics by grounding words, phrases, and complete sentences to objects, their properties, and relations in images. This type of linguistic reasoning is critical for interactions grounded in visually complex environments, such as in robotic applications. However, commonly used resources for language and vision (e.g., Antol et al., 2015; Chen et al., 2016) focus mostly on identification of object properties and few spatial relations (Section 4; Ferraro et al., 2015; Alikhani and Stone, 2019). This relatively simple reasoning, together with biases in the data, removes much of the need to consider language compositionality (Goyal et al., 2017). This motivated the design of datasets that require compositional[1] visual reasoning, including

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 NLVR2. Each caption is paired with two images.[2] The task is to predict if the caption is True or False. The examples require addressing challenging semantic phenomena, including resolving *twice ... as* to counting and comparison of objects, and composing cardinality constraints, such as *at least two dogs in total* and *exactly two*.[3]

NLVR (Suhr et al., 2017) and CLEVR (Johnson et al., 2017a,b). These datasets use synthetic images, synthetic language, or both. The result is a limited representation of linguistic challenges: synthetic languages are inherently of bounded expressivity, and synthetic visual input entails limited lexical and semantic diversity.

We address these limitations with Natural Language Visual Reasoning *for Real* (NLVR2), a new dataset for reasoning about natural language descriptions of photos. The task is to determine if a caption is true with regard to a pair of images. Figure 1 shows examples from NLVR2. We use im-

---

[*] Contributed equally.

[†] Work done as an undergraduate at Cornell University.

[1] In parts of this paper, we use the term *compositional* differently than it is commonly used in linguistics to refer to reasoning that requires composition. This type of reasoning often manifests itself in highly compositional language.
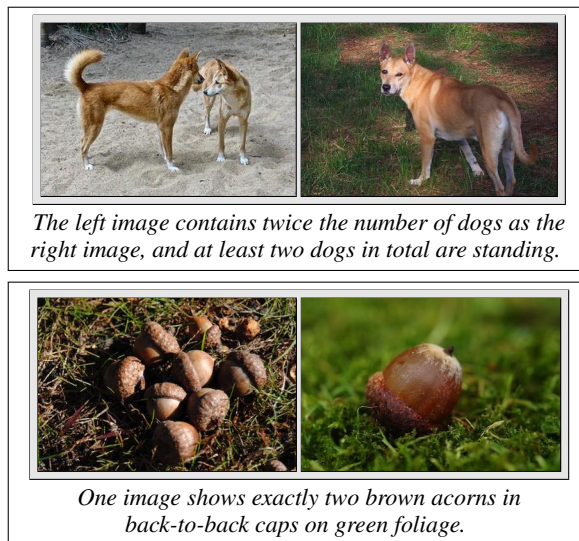
---

[2] Appendix G contains license information for all photographs used in this paper.

[3] The top example is True, while the bottom is False.

ages with rich visual content and a data collection process designed to emphasize semantic diversity, compositionality, and visual reasoning challenges. Our process reduces the chance of unintentional linguistic biases in the dataset, and therefore the ability of expressive models to take advantage of them to solve the task. Analysis of the data shows that the rich visual input supports diverse language, and that the task requires joint reasoning over the two inputs, including about sets, counts, comparisons, and spatial relations.

Scalable curation of semantically-diverse sentences that describe images requires addressing two key challenges. First, we must identify images that are visually diverse enough to support the type of language desired. For example, a photo of a single beetle with a uniform background (Table 2, bottom left) is likely to elicit only relatively simple sentences about the existence of the beetle and its properties. Second, we need a scalable process to collect a large set of captions that demonstrate diverse semantics and visual reasoning.

We use a search engine with queries designed to yield sets of similar, visually complex photographs, including of sets of objects and activities, which display real-world scenes. We annotate the data through a sequence of crowdsourcing tasks, including filtering for interesting images, writing captions, and validating their truth values. To elicit interesting captions, rather than presenting workers with single images, we ask workers for descriptions that compare and contrast four pairs of similar images. The description must be True for two pairs, and False for the other two pairs. Using pairs of images encourages language that composes properties shared between or contrasted among the two images. The four pairs are used to create four examples, each comprising an image pair and the description. This setup ensures that each sentence appears multiple times with both labels, resulting in a balanced dataset robust to linguistic biases, where a sentence's truth value cannot be determined from the sentence alone, and generalization can be measured using multiple image-pair examples.

This paper includes four main contributions: (1) a procedure for collecting visually rich images paired with semantically-diverse language descriptions; (2) NLVR2, which contains 107,292 examples of captions and image pairs, including 29,680 unique sentences and 127,502 im-

ages; (3) a qualitative linguistically-driven data analysis showing that our process achieves a broader representation of linguistic phenomena compared to other resources; and (4) an evaluation with several baselines and state-of-the-art visual reasoning methods on NLVR2. The relatively low performance we observe shows that NLVR2 presents a significant challenge, even for methods that perform well on existing visual reasoning tasks. NLVR2 is available at http://lil.nlp.cornell.edu/nlvr/.

2   Related Work and Datasets

Language understanding in the context of images has been studied within various tasks, including visual question answering (e.g., Zitnick and Parikh, 2013; Antol et al., 2015), caption generation (Chen et al., 2016), referring expression resolution (e.g., Mitchell et al., 2010; Kazemzadeh et al., 2014; Mao et al., 2016), visual entailment (Xie et al., 2019), and binary image selection (Hu et al., 2019). Recently, the relatively simple language and reasoning in existing resources motivated datasets that focus on compositional language, mostly using synthetic data for language and vision (Andreas et al., 2016; Johnson et al., 2017a; Kuhnle and Copestake, 2017; Kahou et al., 2018; Yang et al., 2018).[4] Three exceptions are CLEVR-Humans (Johnson et al., 2017b), which includes human-written paraphrases of generated questions for synthetic images; NLVR (Suhr et al., 2017), which uses human-written captions that compare and contrast sets of synthetic images; and GQA (Hudson and Manning, 2019), which uses synthetic language grounded in real-world photographs. In contrast, we focus on both human-written language and web photographs.

Several methods have been proposed for compositional visual reasoning, including modular neural networks (e.g., Andreas et al., 2016; Johnson et al., 2017b; Perez et al., 2018; Hu et al., 2017; Suarez et al., 2018; Hu et al., 2018; Yao et al., 2018; Yi et al., 2018) and attention- or memory-based methods (e.g., Santoro et al., 2017; Hudson and Manning, 2018; Tan and Bansal, 2018). We use FiLM (Perez et al., 2018), N2NMN (Hu et al., 2017), and MAC (Hudson and Manning, 2018) for our empirical analysis.

In our data, we use each sentence in multiple

---

[4] A tabular summary of the comparison of NLVR2 to existing resources is available in Table 7, Appendix A.

examples, but with different labels. This is related to recent visual question answering datasets that aim to require models to consider both image and question to perform well (Zhang et al., 2016; Goyal et al., 2017; Li et al., 2017; Agrawal et al., 2017, 2018). 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 (Suhr et al., 2017). We adapt this method to web photos, including introducing a process to identify images that support complex reasoning and designing incentives for the more challenging writing task.

3 Data Collection

Each example in NLVR2 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. Our goal is to collect a large corpus of grounded semantically-rich descriptions that require diverse types of reasoning, including about sets, counts, and comparisons. We design a process to identify images that enable such types of reasoning, collect grounded natural language descriptions, and label them as True or False. While we use image pairs, we do not explicitly set the task of describing the differences between the images or identifying which image matches the sentence better (Hu et al., 2019). We use pairs to enable comparisons and set reasoning between the objects that appear in the two images. Figure 2 illustrates our data collection procedure. For further discussion on the design decisions for our task and data collection implementation, please see appendices A and B.

3.1 Image Collection

We require sets of images where the images in each set are detailed but similar enough such that comparison will require use of a diverse set of reasoning skills, more than just object or property identification. Because existing image resources, such as ImageNet (Russakovsky et al., 2015) or COCO (Lin et al., 2014), do not provide such grouping and mostly include relatively simple object-focused scenes, we collect a new set of images. We retrieve sets of images with similar content using search queries generated from synsets from the ILSVRC2014 ImageNet challenge (Russakovsky et al., 2015). This correspon-

dence to ImageNet synsets allows researchers to use pre-trained image featurization models, and focuses the challenges of the task not on object detection, but compositional reasoning challenges.

ImageNet Synsets Correspondence We identify a subset of the 1,000 synsets in ILSVRC2014 that often appear in rich contexts. For example, an `acorn` often appears in images with other `acorns`, while a `seawall` almost always appears alone. For each synset, we issue five queries to the Google Images search engine[5] using query expansion heuristics. The heuristics are designed to retrieve images that support complex reasoning, including images with groups of entities, rich environments, or entities participating in activities. For example, the expansions for the synset `acorn` will include `two acorns` and `acorn fruit`. The heuristics are specified in Table 1. For each query, we use the Google similar images tool for each of the first five images to retrieve the seven non-duplicate most similar images. This results in five sets of eight similar images per query,[6] 25 sets in total. If at least half of the images in a set were labeled as interesting according to the criteria in Table 2, the synset is awarded one point. We choose the 124 synsets with the most points.[7] The 124 synsets are distributed evenly among animals and objects. This annotation was performed by the first two authors and student volunteers, is only used for identifying synsets, and is separate from the image search described below.

Image Search We use the Google Images search engine to find sets of similar images (Figure 2a). We apply the query generation heuristics to the 124 synsets. We use all synonyms in each synset (Deng et al., 2014; Russakovsky et al., 2015). For example, for the synset `timber wolf`, we use the synonym set {`timber wolf`, `grey wolf`, `gray wolf`, `canis lupus` }. For each generated query, we download sets containing at most 16 related images.

Image Pruning We use two crowdsourcing tasks to (1) prune the sets of images, and (2) construct sets of eight images to use in the sentence-writing phase. In the first task, we remove low-

---

[5]https://images.google.com/

[6]At the time of publication, the similar images tool is available at the "View more" link in the list of related images after expanding the results for each image. Images are ranked by similarity, where more similar images appear higher.

[7]We pick 125 and remove one set due to high image pruning rate in later stages.

(a) **Find Sets of Images:** The query `two acorns` 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, similar to this image.


(b) **Image Pruning:** Crowdworkers are given the synset name and identify low-quality images to be removed. In this example, one image is removed because it does not show an instance of the synset `acorn`.

(c) **Set Construction:** Crowdworkers decide whether each of the remaining images is interesting. In this example, three images are marked as non-interesting (top row) because they contain only a single instance of the synset. 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.
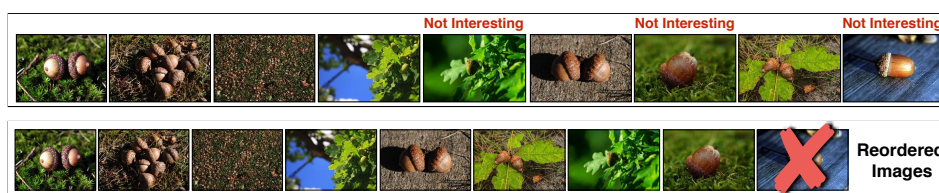
(d) **Sentence Writing:** The images in the set are randomly paired and shown to the worker. The worker selects two pairs, and writes a sentence that is True for the two selected pairs but False for the other two pairs.
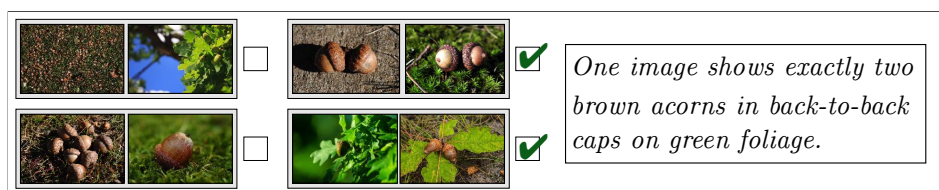
(e) **Validation:** Each pair forms an example with the written sentence. Each example is shown to a worker to re-label.
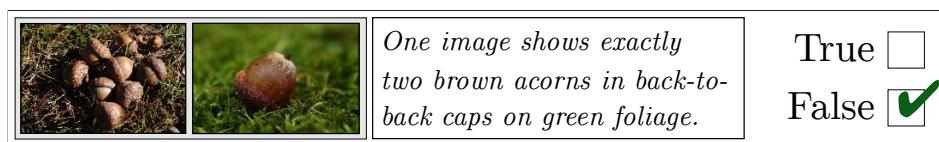

Figure 2: Diagram of the data collection process, showing how a single example from the training set is constructed. Steps (a)–(c) are described in Section 3.1; step (d) in Section 3.2; and step (e) in Section 3.3.

quality images from each downloaded set of similar images (Figure 2b). We display the image set and the synset name, and ask a worker to remove any images that do not load correctly; images that contain inappropriate content, non-realistic artwork, or collages; or images that do not contain an instance of the corresponding synset. This results in sets of sixteen or fewer similar images. We discard all sets with fewer than eight images.

The second task further prunes these sets by removing duplicates and down-ranking non-interesting images (Figure 2c). The goal of this stage is to collect sets that contain enough interesting images. Workers are asked to remove duplicate images, and mark images that are not *in-teresting*. An image is interesting if it fits any of the criteria in Table 2. We ask workers not to mark an image if they consider it interesting for any other reason. We discard sets with fewer than three interesting images. We sort the images in descending order according to first interestingness, and second similarity, and keep the top eight.

3.2 Sentence Writing

Each set of eight images is used for a sentence-writing task. We randomly split the set into four pairs of images. Using pairs encourages comparison and set reasoning within the pairs. Workers are asked to select two of the four pairs and write a sentence that is True for the selected pairs, but

6421

Heuristic	Examples (synset synonym → query)	Description
Quantities	cup → group of cups	Add numerical phrases or manually-identified collective nouns to the synonym. These queries result in images containing multiple examples of the synset.
Hypernyms	flute → flute woodwind	Add direct or indirect hypernyms from WordNet (Miller, 1993). Applied only to the non-animal synsets. This heuristic increases the diversity of images retrieved for the synset (Deng et al., 2014).
Similar words	banana → banana pear	Add concrete nouns whose cosine similarity with the synonym is greater than 0.35 in the embedding space of Google News word2vec embeddings (Mikolov et al., 2013). Applied only to non-animal synsets. These queries result in images containing a variety of different but related object types.
Activities	beagle → beagles eating	Add manually-identified verbs describing common activities of animal synsets. Applied only to animal synsets. This heuristic results in images of animals participating in activities, which encourages captions with a diversity of entity properties.

Table 1: The four heuristics used to generate search queries from synsets.

Positive Examples and Criteria

Contains more than one instance of the synset.

Shows an instance of the synset interacting with other objects.

Shows an instance of the synset performing an activity.

Displays a set of diverse objects or features.

Negative Examples

Table 2: Positive and negative examples of interesting images.

False for the unselected pairs. Allowing workers to select pairs themselves makes the sentence-writing task easier than with random selection, which may create tasks that are impossible to complete. Writing requires finding similarities and differences between the pairs, which encourages compositional language (Suhr et al., 2017).

In contrast to the collection process of NLVR, using real images does not allow for as much control over their content, in some cases permitting workers to write simple sentences. For example, a worker could write a sentence stating the existence of a single object if it was only present in both selected pairs, which is avoided in NLVR by controlling for the objects in the images. Instead, we define more specific guidelines for the workers for writing sentences, including asking to avoid subjective opinions, discussion of properties of photograph, mentions of text, and simple object identification. We include more details and examples of these guidelines in Appendix B.

3.3 Validation

We split each sentence-writing task into four examples, where the sentence is paired with each pair of images. Validation ensures that the selection of each image pair reflects its truth value. We show each example independently to a worker, and ask them to label it as True or False. The worker may also report the sentence as nonsensical. We keep all non-reported examples where the validation label is the same as the initial label indicated by the sentence-writer's selection. For example, if the image pair is initially selected during sentence-writing, the sentence-writer intends the sentence to be True for the pair, so if the validation label is False, this example is removed.

3.4 Splitting the Dataset

We assign a random 20% of the examples passing validation to development and testing, ensuring that examples from the same initial set of eight images do not appear across the split. For these examples, we collect four additional validation judgments to estimate agreement and human performance. We remove from this set examples where two or more of the extra judgments disagreed with the existing label (Section 3.3). Finally, we create

One image contains a single vulture in a standing pose with its head and body facing leftward, and the other image contains a group of at least eight vultures.

There are two trains in total traveling in the same direction.

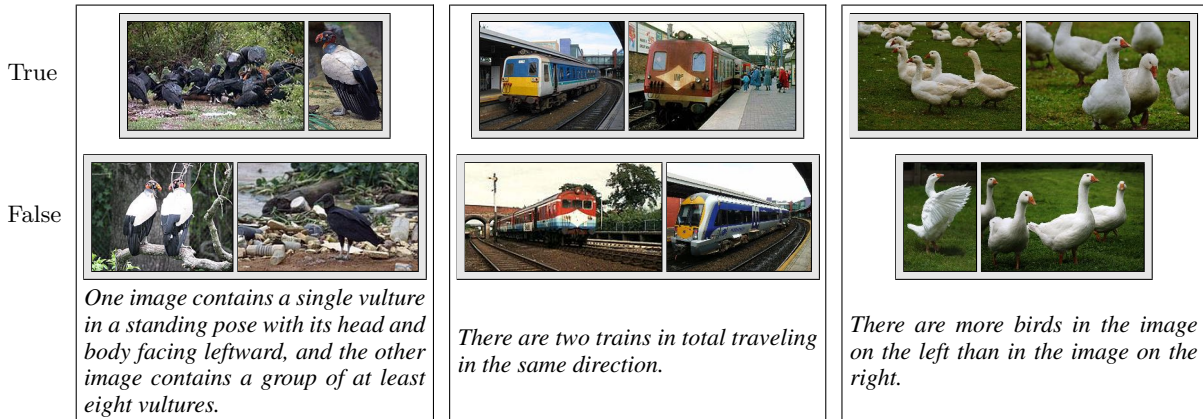*There are more birds in the image on the left than in the image on the right.*

Table 3: Six examples with three different sentences from NLVR2. For each sentence, we show two examples using different image-pairs, each with a different label.

equal-sized splits for a development set and two test sets, ensuring that original image sets do not appear in multiple splits of the data (Table 4).

3.5 Data Collection Management

We use a tiered system with bonuses to encourage workers to write linguistically diverse sentences. After every round of annotation, we sample examples for each worker and give bonuses to workers that follow our writing guidelines well. Once workers perform at a sufficient level, we allow them access to a larger pool of tasks. We also use qualification tasks to train workers. The mean cost per unique sentence in our dataset is $0.65; the mean cost per example is $0.18. Appendix B provides additional details about our bonus system, qualification tasks, and costs.

3.6 Collection Statistics

We collect 27,678 sets of related images and a total of 387,426 images (Section 3.1). Pruning low-quality images leaves 19,500 sets and 250,862 images. Most images are removed for not containing an instance of the corresponding synset or for being non-realistic artwork or a collage of images. We construct 17,685 sets of eight images each.

We crowdsource 31,418 sentences (Section 3.2). We create two writing tasks for each set of eight images. Workers may flag sets of images if they should have been removed in earlier stages; for example, if they contain duplicate images. Sentence-writing tasks that remain without annotation after three days are removed.

During validation, 1,875 sentences are reported as nonsensical. 108,516 examples pass validation; i.e., the validation label matches the initial selec-

	Unique sentences	Examples
Train	23,671	86,373
Development	2,018	6,982
Test-P	1,995	6,967
Test-U	1,996	6,970
Total	29,680	107,292

Table 4: NLVR2 data splits.

tion for the pair of images (Section 3.3). Removing low-agreement examples in the development and test sets yields a dataset of 107,292 examples, 127,502 unique images, and 29,680 unique sentences. Each unique sentence is paired with an average of 3.6 pairs of images. Table 3 shows examples of three unique sentences from NLVR2. Table 4 shows the sizes of the data splits, including train, development, a public test set (Test-P), and an unreleased test set (Test-U).

4 Data Analysis

We perform quantitative and qualitative analysis using the training and development sets.

Agreement Following validation, 8.5% of the examples not reported during validation are removed due to disagreement between the validator's label and the initial selection of the image pair (Section 3.3).[8] We use the five validation labels we collect for the development and test sets to compute Krippendorff's α and Fleiss' κ to measure agreement (Cocos et al., 2015; Suhr et al., 2017). Before removing low-agreement examples

---

[8]The validator is the same worker as the sentence-writer for 11.5% of examples. In these cases, the validator agrees with themselves 96.7% of the time. For examples where the sentence-writer and validator were not the same person, they agree in 90.8% of examples.
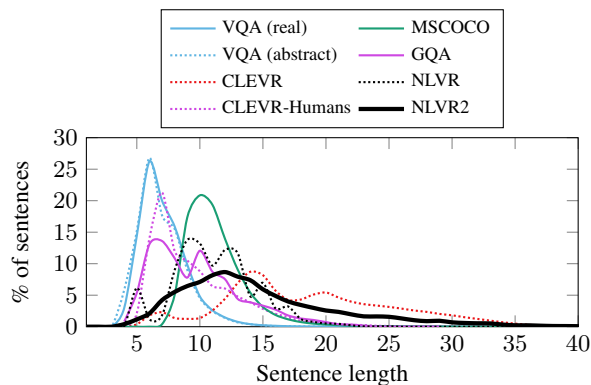
Figure 3: Distribution of sentence lengths. Dotted curves represent datasets with synthetic images.

(Section 3.4), $\alpha = 0.906$ and $\kappa = 0.814$. After removal, $\alpha = 0.912$ and $\kappa = 0.889$, indicating almost perfect agreement (Landis and Koch, 1977).

Synsets Each synset is associated with $\mu = 752.9 \pm 205.7$ examples. The five most common synsets are `gorilla`, `bookcase`, `bookshop`, `pug`, and `water buffalo`. The five least common synsets are `orange`, `acorn`, `ox`, `dining table`, and `skunk`. Synsets appear in equal proportions across the four splits.

Language NLVR2's vocabulary contains 7,457 word types, significantly larger than NLVR, which has 262 word types. Sentences in NLVR2 are on average 14.8 tokens long, whereas NLVR has a mean sentence length of 11.2. Figure 3 shows the distribution of sentence lengths compared to related corpora. NLVR2 shows a similar distribution to NLVR, but with a longer tail. NLVR2 contains longer sentences than the questions of VQA (Antol et al., 2015), GQA (Hudson and Manning, 2019), and CLEVR-Humans (Johnson et al., 2017b). Its distribution is similar to MSCOCO (Chen et al., 2015), which also contains captions, and CLEVR (Johnson et al., 2017a), where the language is synthetically generated.

We analyze 800 sentences from the development set for occurrences of semantic and syntactic phenomena (Table 5). We compare with the 200-example analysis of VQA and NLVR from Suhr et al. (2017), and 200 examples from the balanced split of GQA. Generally, NLVR2 has similar linguistic diversity to NLVR, showing broader representation of linguistic phenomena than VQA and GQA. One noticeable difference from NLVR is less use of hard cardinality. This is possibly due to how NLVR is designed to use a very limited set

of object attributes, which encourages writers to rely on accurate counting for discrimination more often. We include further analysis in Appendix C.

5 Estimating Human Performance

We use the additional labels of the development and test examples to estimate human performance. We group these labels according to workers. We do not consider cases where the worker labels a sentence written by themselves. For each worker, we measure their performance as the proportion of their judgements that matches the gold-standard label, which is the original validation label. We compute the average and standard deviation performance over workers with at least 100 such additional validation judgments, a total of 68 unique workers. Before pruning low-agreement examples (Section 3.4), the average performance over workers in the development and both test sets is 93.1 ± 3.1. After pruning, it increases to 96.1 ± 2.6. Table 6 shows human performance for each data split that has extra validations. Because this process does not include the full dataset for each worker, it is not fully comparable to our evaluation results. However, it provides an estimate by balancing between averaging over many workers and having enough samples for each worker.

6 Evaluation Systems

We evaluate several baselines and existing visual reasoning approaches using NLVR2. For all systems, we optimize for example-level accuracy.[9]

We measure the biases in the data using three baselines: (a) MAJORITY: assign the most common label (True) to each example; (b) TEXT: encode the caption using a recurrent neural network (RNN; Elman, 1990), and use a multilayer perceptron to predict the truth value; and (c) IMAGE: encode the pair of images using a convolutional neural network (CNN), and use a multilayer perceptron to predict the truth value. The latter two estimate the potential of solving the task using only one of the two modalities.

We use two baselines that consider both language and vision inputs. The CNN+RNN baseline concatenates the encoding of the text and images, computed similar to the TEXT and IMAGE baselines, and applies a multilayer perceptron to predict a truth value. The MAXENT baseline computes features from the sentence and objects de-

---

[9]System and learning details are available in Appendix E.

	VQA (real) %	GQA %	NLVR %	NLVR2 %	Example from NLVR2
Semantics					
Cardinality (hard)	11.5	0	66	41.1	*Six rolls of paper towels are enclosed in a plastic package with the brand name on it.*
Cardinality (soft)	1	0	23.6	22.5	*No more than two cheetahs are present.*
Existential	11.5	16.5	88	23.6	*There are at most 3 water buffalos in the image pair.*
Universal	1	4.5	7.5	16.8	*In one image there is a line of fence posts with one large darkly colored bird on top of each post.*
Coordination	5	21.5	17	33.3	*Each image contains only one wolf, and all images include snowy backdrops.*
Coreference	6.5	0.5	3	14.6	*there are four or more animals very close to each other on the grass in the image to the left.*
Spatial Relations	42.5	43	66	49	*A stylus is near a laptop in one of the images.*
Comparative	1	2	3	8	*There are more birds in the image on the right than in the image on the left.*
Presupposition	80	79	19.5	20.6	*A cookie sits in the dessert in the image on the left.*
Negation	1	2.5	9.5	9.6	*The front paws of the dog in the image on the left are not touching the ground.*
Syntactic Ambiguity					
CC Attachment	0	2.5	4.5	3.8	*The left image shows a cream-layered dessert in a footed clear glass which includes sliced peanut butter cups and brownie chunks.*
PP Attachment	3	6.5	23	11.5	*At least one panda is sitting near a fallen branch on the ground.*
SBAR Attachment	0	5	2	1.9	*Balloons float in a blue sky with dappled clouds on strings that angle rightward, in the right image.*

Table 5: Linguistic analysis of sentences from NLVR2, GQA, VQA, and NLVR. We analyze 800 development sentences from NLVR2 and 200 from each of the other datasets for the presence of semantic and syntactic phenomena described in Suhr et al. (2017). We report the proportion of examples containing each phenomenon.

tected in the paired images. We detect the objects in the images using a Mask R-CNN model (He et al., 2017; Girshick et al., 2018) pre-trained on the COCO detection task (Lin et al., 2014). We use a detection threshold of 0.5. For each n-gram with a numerical phrase in the caption and object class detected in the images, we compute features based on the number present in the n-gram and the detected object count. We create features for each image and for both together, and use these features in a maximum entropy classifier.

Several recent approaches to visual reasoning make use of modular networks (Section 2). Broadly speaking, these approaches predict a neural network layout from the input sentence by using a set of modules. The network is used to reason about the image and text. The layout predictor may be trained: (a) using the formal programs used to generate synthetic sentences (e.g., in CLEVR), (b) using heuristically generated layouts from syntactic structures, or (c) jointly with the neural modules with latent layouts. Because sentences in NLVR2 are human-written, no supervised formal programs are available at training time. We use two methods that do not require

such formal programs: end-to-end neural module networks (N2NMN; Hu et al., 2017) and feature-wise linear modulation (FiLM; Perez et al., 2018). For N2NMN, we evaluate three learning methods: (a) N2NMN-CLONING: using supervised learning with gold layouts; (b) N2NMN-TUNE: using policy search after cloning; and (c) N2NMN-RL: using policy search from scratch. For N2NMN-CLONING, we construct layouts from constituency trees (Cirik et al., 2018). Finally, we evaluate the Memory, Attention, and Composition approach (MAC; Hudson and Manning, 2018), which uses a sequence of attention-based steps. We modify N2NMN, FiLM, and MAC to process a pair of images by extracting image features from the concatenation of the pair.

7 Experiments and Results

We use two metrics: accuracy and consistency. Accuracy measures the per-example prediction accuracy. Consistency measures the proportion of unique sentences for which predictions are correct for all paired images (Goldman et al., 2018). For training and development results, we report mean and standard deviation of accuracy and con-

	Train	Dev	Test-P	Test-U
MAJORITY (assign True)	50.8/2.1	50.9/3.9	51.1/4.2	51.4/4.6
TEXT	50.8±0.0/2.1±0.0	50.9±0.0/3.9±0.0	51.1/4.2	51.4/4.6
IMAGE	60.1±2.9/14.2±4.2	51.6±0.2/8.4±0.8	51.9/7.4	51.9/7.1
CNN+RNN	94.3±3.3/84.5±10.2	53.4±0.4/12.2±0.7	52.4/11.0	53.2/11.2
MAXENT	89.4/73.4	54.1/11.4	**54.8/11.5**	**53.5/12.0**
N2NMN (Hu et al., 2017):				
N2NMN-CLONING	65.7±25.8/30.8±49.7	50.2±1.0/5.7±3.1	–	–
N2NMN-TUNE	96.5±1.6/94.9±0.4	50.0±0.7/9.8±0.5	–	–
N2NMN-RL	50.8±0.3/2.3±0.3	51.0±0.1/4.1±0.3	51.1/5.0	51.5/5.0
FiLM (Perez et al., 2018)	69.0±16.9/32.4±29.6	51.0±0.4/10.3±1.0	52.1/9.8	53.0/10.6
MAC (Hudson and Manning, 2018)	87.4±0.8/64.0±1.7	50.8±0.6/11.0±0.2	51.4/11.4	51.2/11.2
HUMAN	–	96.2±2.1/−	96.3±2.9/−	96.1±3.1/−

Table 6: Performance (accuracy/consistency) on NLVR2.

sistency over three trials as $\mu_{\text{acc}} \pm \sigma_{\text{acc}} / \mu_{\text{cons}} \pm \sigma_{\text{cons}}$. The results on the test sets are generated by evaluating the model that achieved the highest accuracy on the development set. For the N2NMN methods, we report test results only for the best of the three variants on the development set.[10]

Table 6 shows results for NLVR2. MAJORITY results demonstrate the data is fairly balanced. The results are slightly higher than perfect balance due to pruning (Sections 3.3 and 3.4). The TEXT and IMAGE baselines perform similar to MAJORITY, showing that both modalities are required to solve the task. TEXT shows identical performance to MAJORITY because of how the data is balanced. The best performing system is the feature-based MAXENT with the highest accuracy and consistency. FiLM performs best of the visual reasoning methods. Both FiLM and MAC show relatively high consistency. While almost all visual reasoning methods are able to fit the data, an indication of their high learning capacity, all generalize poorly. An exception is N2NMN-RL, which fails to fit the data, most likely due to the difficult task of policy learning from scratch. We also experimented with recent contextualized word embeddings to study the potential of stronger language models. We used a 12-layer uncased pre-trained BERT model (Devlin et al., 2019) with FiLM. We observed BERT provides no benefit, and therefore use the default embedding method for each model.

8 Conclusion

We introduce the NLVR2 corpus for studying semantically-rich joint reasoning about photographs and natural language captions. Our focus on visually complex, natural photographs and human-written captions aims to reflect the challenges of compositional visual reasoning better than existing corpora. Our analysis shows that the language contains a wide range of linguistic phenomena including numerical expressions, quantifiers, coreference, and negation. This demonstrates how our focus on complex visual stimuli and data collection procedure result in compositional and diverse language. We experiment with baseline approaches and several methods for visual reasoning, which result in relatively low performance on NLVR2. These results and our analysis exemplify the challenge that NLVR2 introduces to methods for visual reasoning. We release training, development, and public test sets, and provide scripts to break down performance on the 800 examples we manually analyzed (Section 4) according to the analysis categories. Procedures for evaluating on the unreleased test set and a leaderboard are available at `http://lic.nlp.cornell.edu/nlvr/`.

---

[10]For reference, we also provide NLVR results in Table 11, Appendix D.

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4971–4980.

Aishwarya Agrawal, Aniruddha Kembhavi, Dhruv Batra, and Devi Parikh. 2017. C-VQA: A compositional split of the visual question answering (VQA) v1.0 dataset. *CoRR*, abs/1704.08243.

Malihe Alikhani and Matthew Stone. 2019. "Caption" as a coherence relation: Evidence and implications. In *Proceedings of the Workshop on Shortcomings in Vision and Language*.

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

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

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

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325.

Volkan Cirik, Taylor Berg-Kirkpatrick, and Louis-Philippe Morency. 2018. Using syntax to ground referring expressions in natural images. In *AAAI Conference on Artificial Intelligence*.

Anne Cocos, Aaron Masino, Ting Qian, Ellie Pavlick, and Chris Callison-Burch. 2015. Effectively crowdsourcing radiology report annotations. In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*, pages 109–114.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive Science*, 14:179–211.

Francis Ferraro, Nasrin Mostafazadeh, Ting-Hao Huang, Lucy Vanderwende, Jacob Devlin, Michel Galley, and Margaret Mitchell. 2015. A survey of current datasets for vision and language research. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 207–213.

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

Omer Goldman, Veronica Latcinnik, Ehud Nave, Amir Globerson, and Jonathan Berant. 2018. Weakly supervised semantic parsing with abstract examples. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1809–1819.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6325–6334.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. 2017. Mask R-CNN. In *IEEE International Conference on Computer Vision*, pages 2980–2988.

Hexiang Hu, Ishan Misra, and Laurens van der Maaten. 2019. Binary image selection (BISON): Interpretable evaluation of visual grounding. *CoRR*, abs/1901.06595.

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

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

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

Drew A. Hudson and Christopher D. Manning. 2019. GQA: a new dataset for compositional question answering over real-world images. In *IEEE Conference on Computer Vision and Pattern Recognition*.

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

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

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

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 787–798.

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

J. Richard Landis and Gary Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33 1:159–74.

Yining Li, Chen Huang, Xiaoou Tang, and Chen Change Loy. 2017. Learning to disambiguate by asking discriminative questions. In *IEEE International Conference on Computer Vision*, pages 3439–3448.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11–20.

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

George A. Miller. 1993. WordNet: A lexical database for English. In *Proceedings of the Workshop on Human Language Technology*, pages 409–409.

Margaret Mitchell, Kees van Deemter, and Ehud Reiter. 2010. Natural reference to objects in a visual domain. In *Proceedings of the International Natural Language Generation Conference*.

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

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. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252.

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

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

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A corpus of natural language for visual reasoning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 217–223.

Hao Tan and Mohit Bansal. 2018. 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.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *CoRR*, abs/1901.06706.

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

Yiqun Yao, Jiaming Xu, Feng Wang, and Bo Xu. 2018. Cascaded mutual modulation for visual reasoning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 975–980. Association for Computational Linguistics.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. 2018. Neural-symbolic VQA: Disentangling reasoning from vision and language understanding. In *Advances in Neural Information Processing Systems*, pages 1031–1042.

Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Yin and yang: Balancing and answering binary visual questions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5014–5022.

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