A Corpus for Reasoning About Natural Language Grounded in Photographs (Appendix)

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

A Frequently Asked Questions

In what applications do you expect to see the kind of language NLVR2 allows to study? Composition of reasoning skills including counting, comparing, and reasoning about sets is critical for robotic agents following natural language instructions. Consider a robot on a factory floor or in a cluttered workshop following the instruction *get the two largest hammers from the toolbox at the end of the shelf*. Correctly following this instruction requires reasoning compositionally about object properties, comparisons between these properties, counts of objects, and spatial relations between observed objects. The language in NLVR2 reflects this type of linguistic reasoning. While the task we define does not use this kind of application directly, our data enables studying models that can understand this type of language.

How can I use NLVR2 to build an end application? The task and data are not intended to directly develop an end application. Our focus is on developing a task that drives research in vision and language understanding towards handling diverse set of reasoning skills. It is critical to keep in mind that this dataset was not analyzed for social biases. Researchers who wish to apply this work to an end product should take great care in considering what biases may exist.

Doesn't using a binary prediction task limit the ability to gain insight into model performance? Because our dataset contains both positive and negative image pairs for each sentence, we can measure consistency (Goldman et al., 2018), which requires a model to predict each label correctly for each use of the sentence. This metric requires generalization across at most four image pair contexts.

Why collect a new set of images rather than use existing ones like COCO (Lin et al., 2014)? Our goal was to achieve similar semantic diversity to NLVR, but using real images. Like NLVR, we use a sentence-writing task where sets of similar images are compared and contrasted. However, unlike NLVR, we do not have control over the image content, so cannot guarantee image sets where the content is similar enough (e.g., where the only difference is the direction in which the same animal is facing) such that the written sentence does not describe trivial image differences (e.g., the types of objects present). In addition to image similarity within sets, we also prioritize image interestingness, for example images with many instances of an object. Existing corpora, including like COCO and ImageNet (Russakovsky et al., 2015), were not constructed to prioritize interestingness as we define it, and are not comprised of sets of eight very similar images as required for our task.

1. We select a set of 124 ImageNet synsets which often appear in visually rich images.

2. We generate search queries which result in visually rich images, e.g., containing multiple instances of a synset.

3. We use a similar images tool to acquire sets of images with similar image content, for example containing the same objects in different relative orientations.

4. We prune images which do not contain an example of the synset it was derived from.

5. We apply a re-ranking and pruning procedure that prioritizes visually rich and interesting images, and prunes set which do not have enough interesting images.

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

These steps result in a total of 17,685 sets of eight similar, visually rich images.

Why use pairs of images instead of single images? We use pairs of images to elicit descriptions that reason over the pair of images in addition to the content within each image. This setup supports, for example, comparing the two images, requiring that a condition holds in both images or in one but not the other, and performing set reasoning about the objects present in each image. This is analogous to the three-box setup in NLVR.

Why allow workers to select the pairs themselves during sentence writing? We found that for some image pair selections, it was too difficult for workers to write a sentence which distinguishes the pairs. Allowing the workers to choose the pairs avoids this feasiblity issue.

Why get multiple validations for development and test splits? This ensures the test splits are of the highest quality and have minimal noise, as required for reliable measure of task performance. The additional annotatiosn also allow us to measure agreement and estimate human performance.

How does the NLVR2 data compares to the NLVR data? NLVR and NLVR2 share the task of determining whether a sentence is true in a given visual context. In NLVR, the visual input is synthetic and includes a handful of shapes and properties. In NLVR2, each visual context is a pair of real photographs obtained from the web. Grounding sentences in image pairs rather than single images is related to NLVR's use of three boxes per image.

How does the NLVR2 data collection process compare to NLVR? We adapt the NLVR sentence-writing and validation tasks. However, rather than using four related synthetic images for writing, we use four pairs of real images. The pairing of images encourages set comparison. This was accomplished in NLVR through careful control of the generated image content, something that is not possible with real images. The NLVR image generation process is also controlled for the type of differences possible between images and the visual complexity, by ensuring the objects present in the selected and unselected images were the same. This guarantees that the only differences are in the object configurations and distribution among the three boxes in each image. Neither form of control is possible with real images. Instead, we rewrite the guidelines and develop a process to educate workers to follow them. In our process, we use the similar images tool to identify images that require linguistically-rich descriptions to distinguish. While using the similar images tool does not guarantee that the objects in the selected images are also present in the unselected images, our process successfully avoids this issue; in practice, only around 13% of examples take advantage of this by mentioning objects only present in the selected images.

Can you summarize the key linguistic differences between NLVR2 and NLVR? NLVR contains significantly[1] more examples of hard cardinality, existential quantifiers, spatial relations, and prepositional attachment ambiguity. NLVR2 contains significantly[1] more examples of universal quantifiers, coordination, coreference, and comparatives. NLVR2's descriptions are longer on average than NLVR (14.8 vs. 11.2 tokens), and the vocabulary is much larger (7,457 vs. 262 word types). This demonstrates both the lexical diversity and challenges of understanding a wide range of image content in NLVR2 that are not present in NLVR. However, NLVR allows studying compositionality in isolation from lexical diversity, an intended feature of the dataset's design. NLVR has also been used as a semantic parsing task, where images are represented as structured representations (Goldman et al., 2018), a use case that is not possible with NLVR2. NLVR remains a challenging dataset for visual reasoning; recent approaches have shown moderate improvements over the initial baseline performance, yet remain far from human accuracy, which we compute in Table 11.

How does NLVR2 compare to existing visual reasoning datasets? Table 7 compares NLVR2 with several existing, related corpora. In the last several years there has been an increase in the number of datasets released for vision and language research. One trend includes building datasets for compositional visual reasoning (SHAPES, CLEVR, CLEVR-Humans, ShapeWorld, NLVR, FigureQA, COG, and GQA), all of which use synthetic data either for at least one of the inputs. While NLVR2 requires related visual reasoning skills, it uses both real natural language and real visual inputs.

How does NLVR2 compare to recent attempts to avoid biases in vision and language datasets? Recently, several approaches were proposed to

[1] Using a χ^2 test with $p < 0.05$.

identify unintended biases present in vision-and-language tasks, such as the ability to answer a question without using the paired image (Zhang et al., 2016; Goyal et al., 2017; Li et al., 2017; Agrawal et al., 2017, 2018). The data collection process of NLVR2 is designed to automatically pair each sentence with both labels in different visual contexts. This makes NLVR2 robust to implicit linguistic biases. This is illustrated by our initial experiments with BERT, which have been shown to be extremely effective at capturing language patterns for various tasks (Devlin et al., 2019). With our balanced data, using BERT does not help identifying and using language biases.

Are the differences in the linguistic analysis between the datasets significant? We measure significance using a χ^2 test with $p < 0.05$. Our qualitative linguistic analysis shows several differences from VQA (Antol et al., 2015) and GQA (Hudson and Manning, 2019). NLVR2 contains significantly more examples of hard cardinality, soft cardinality, existential quantifiers, universal quantifiers, coordination, coreference, spatial relations, comparatives, negation, and preposition attachment ambiguity than both GQA and VQA. However, VQA and GQA both contain significantly more examples of presupposition than NLVR2.

Given your linguistic analysis, how does GQA compare to VQA? We found that the distribution of phenomena in VQA and GQA are roughly similar, with notable differences being significantly[1] more examples of hard cardinality and coreference in VQA, and significantly[1] more examples of universal quantifiers, coordination, and coordination and subordinating conjunction attachment ambiguity in GQA.

B Data Collection Details

Image Collection We consider the images of each search query in the order of the search results. For each result associated with a set of similar images, we save the URL of the result image and the URLs of the fifteen most similar images, giving us a set of sixteen images. We skip and ignore URLs from a hand-crafted list of stock photo domains; images from these domains include large, distracting watermarks. We stop after observing 60 result images, saving 30 sets of image URLs, or observing five consecutive results that do not have similar images.[2]

After downloading a set of 16 URLs of related images (Section 3.1), we automatically prune the images. We remove any broken URLs or any URLS that appeared in other previously-downloaded sets from the same search query. We remove downloaded images smaller than 200×200 pixels. We apply basic duplicate removal by removing any images which are exact duplicates of a previously-downloaded image in the set. This automatic pruning may result in image sets consisting of fewer than 16 images. We discard any sets after this stage with fewer than 8 images.

Sentence Writing Table 8 shows the types of sentences we ask workers to avoid in their writing. Analysis of 100 sentences from the development set shows that almost all sentences follow our guidelines, only 13% violate our guidelines. The most common violation was mentioning an object not present in the unselected images. Such sentences can trivially be labeled as False in the context of the unselected pairs, as the mentioned object will not be present. In the context of the selected pairs, however, a model must still perform compositional joint reasoning about the sentence and the image pair to determine whether the label should be True at test time. This is because the sentence often includes additional constraints. The bottom example in Table 12 illustrates this violation. A system may easily determine that because neither a hole nor a golf flagpole are present in either image, the sentence is False. However, if these objects were present, the system must reason about counts and spatial relations of the mentioned objects to verify that the sentence is True.

Data Collection Management We use two qualification tasks. For the set construction and sentence writing tasks, we qualify workers by first showing six tutorial questions about the guidelines and task. We then ask them to validate guidelines for nineteen sentences across two sets of four pre-selected image pairs, and to complete a single sentence-writing task for pre-selected image pairs. We validate the written sentence by hand. We qualify workers for validation with eight pre-selected validation tasks.

We use a bonus system to encourage workers to write linguistically diverse sentences. We conduct sentence writing in rounds. After each round,

[2] For collective nouns and the numerical phrase `two <synset>`, we instead observe at most 100 top images or save at most 60 sets.

Dataset	Task	Prevalent Linguistic Phenomena	Natural Language?	Natural Images?
NLVR2	Binary Sentence Classification	(1) Hard and (2) soft cardinality; (3) existential and (4) universal quantifiers; (5) coordination; (6) coreference; (7) spatial relations; (8) presupposition; (9) preposition attachment ambiguity	✔	✔
VQA1.0 (Antol et al., 2015), VQA-CP (Agrawal et al., 2017), VQA2.0 (Goyal et al., 2017)	Visual Question Answering	(1) Hard cardinality; (2) existential quantifiers; (3) spatial relations; (4) presupposition	✔	✔
NLVR (Suhr et al., 2017)	Binary Sentence Classification	(1) Hard and (2) soft cardinality; (3) existential quantifiers; (4) coordination; (5) spatial relations; (6) presupposition; (7) preposition attachment ambiguity	✔	
GQA (Hudson and Manning, 2019)	Visual Question Answering	(1) Existential quantifiers; (2) coordination; (3) spatial relations; (4) presupposition		✔

Dataset	Task	Natural Language?	Natural Images?
SAIL (MacMahon et al., 2006)	Instruction Following	✔	
Mitchell et al. (2010)	Referring Expression Resolution	✔	
Matuszek et al. (2012)	Referring Expression Resolution	✔	
FitzGerald et al. (2013)	Referring Expression Generation	✔	
VQA (Abstract) (Zitnick and Parikh, 2013)	Visual Question Answering	✔	
ReferItGame (Kazemzadeh et al., 2014)	Referring Expression Resolution	✔	✔
SHAPES (Andreas et al., 2016)	Visual Question Answering		
Bisk et al. (2016)	Instruction Following	✔	
MSCOCO (Chen et al., 2016)	Caption Generation	✔	✔
Google RefExp (Mao et al., 2016)	Referring Expression Resolution	✔	✔
ROOM-TO-ROOM (Anderson et al., 2018)	Instruction Following	✔	✔
Visual Dialog (Das et al., 2017)	Dialogue Visual Question Answering	✔	✔
CLEVR (Johnson et al., 2017a)	Visual Question Answering		
CLEVR-Humans (Johnson et al., 2017b)	Visual Question Answering	✔	
TDIUC (Kafle and Kanan, 2017)	Visual Question Answering	✔	✔
ShapeWorld (Kuhnle and Copestake, 2017)	Binary Sentence Classification		
FigureQA (Kahou et al., 2018)	Visual Question Answering		
TVQA (Lei et al., 2018)	Video Question Answering	✔	✔
LANI & CHAI (Misra et al., 2018)	Instruction Following	✔	✔
Talk the Walk (de Vries et al., 2018)	Dialogue Instruction Following	✔	✔
COG (Yang et al., 2018)	Visual Question Answering; Instruction Following		
VCR (Zellers et al., 2019)	Visual Question Answering	✔	✔
TallyQA (Acharya et al., 2019)	Visual Question Answering	✔	✔
TOUCHDOWN (Chen et al., 2019)	Instruction Following; Spatial Description Resolution	✔	✔
COCO-BISON (Hu et al., 2019)	Binary Image Selection	✔	✔
SNLI-VE (Xie et al., 2019)	Visual Entailment	✔	✔

Table 7: Comparison between NLVR2 and existing datasets for language and vision research. The top table details prevalent linguistic phenomena in some of the most related datasets according to our analysis, listing each linguistic phenomenon with at least 10% representation as prevalent. For each dataset, we count the number of prevalent phenomena. NLVR2 has the broadest representation. The bottom table lists other tasks in language and vision.

What to avoid	Example of erroneous sentence
Subjective opinions	*The dog's fur has a nice color pattern.*
Discussing properties of the photograph	*In both images, the cat's paw is cropped out of the photo.*
Mentioning text in the photograph	*Both trains are numbered 72.*
Mentioned object not present in unselected pairs	*There is a cup on top of a chair.* – for a set of images where the selected pairs contain a chair, but the unselected pairs do not.
Mentioning the presence of a single object	*There is a hammer.*
Disjunction on images in the pair	*The left image contains a penguin, and the right image contains a rock.*

Table 8: Types of sentences workers are discouraged from writing. The bottom two are permissible as long as the sentence includes other kinds of reasoning.

	Cost	Unique Workers
Image Pruning	$1,310.76	53
Set Construction	$1,798.84	46
Sentence Writing	$9,570.46	99
Validation	$6,452.93	125
Total	$19,132.99	167

Table 9: Cost and worker statistics.

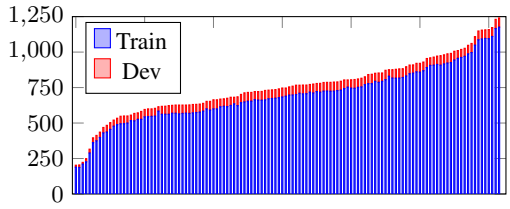


Figure 4: Number of examples per synset, sorted by number of examples in each synset.

we sample twenty sentences for each worker from that round. If at least 75% of these sentences follow the guidelines, they receive a bonus for each sentence written during the last round. If between 50% and 75% follow our guidelines, they receive a slightly lower bonus. This encourages workers to follow the guidelines more closely. In addition, each worker initially only has access to a limited pool of sentence-writing tasks. Once they successfully complete an evaluation round where at least 75% of their sentences followed the guidelines, they get access to the entire pool of tasks.

Table 9 shows the costs and number of workers per task. The final cost per unique sentence in our dataset is $0.65; the cost per example is $0.18.

C Additional Data Analysis

Synsets Figure 4 shows the counts of examples per synset in the training and development sets.

Image Pair Reasoning We use a 200-sentence subset of the sentences analyzed in Table 5 to analyze what types of reasoning are required over the two images (Table 10). We observe that sentences commonly use the pair structure used to display the images: 11% of sentences require that a property to hold in both images, 19% simply require that a property holds in at least one image, and 26.5% of sentences require a property to be true in the left or right images specifically. The pair is also used for comparison, with 6% of sentences requiring comparing properties of the two images. Finally, 39.5% of sentences simply state a property that must be true across the image pair, e.g., *One sliding door is closed.*

D Results on NLVR

Table 11 shows previously published results using raw images in NLVR from Suhr et al. (2017) and more recent approaches.[3] We also report results for visual reasoning systems originally developed for CLEVR. We compute human performance for each split of the data using the procedure described in Section 5; a threshold of 100 covers 100% of annotators. NMN (Andreas et al., 2016), N2NMN, and FiLM achieve the best results for methods that were not developed using NLVR. However, both perform worse than CNN-BIATT (Tan and Bansal, 2018) and CMM (Yao et al., 2018), which were developed originally using NLVR.[4]

E Implementation Details

For the TEXT, IMAGE, and CNN+RNN baselines, we first compute a representation of the input(s). We then process this representation using a multilayer perceptron (MLP). The MLP's output is used to predict a distribution over the two labels using a softmax. The MLP includes learned bias terms and ReLu nonlinearities on the output of each layer, except the last one. In all cases, the layer sizes of the MLP follow the

[3] Not all previously evaluated methods report consistency.

[4] Consistency for CNN-BIATT was taken from the NLVR leaderboard.

Required Reasoning	%	Example from NLVR2
Exactly one image	3	***Only one image*** *shows warthogs butting heads.*
Existential quantification	19	***In one image****, hyenas fight with a big cat.*
Universal quantification	11	*There are people walking* ***in both images****.*
Explicit reference to left and/or right image	26.5	***The left image*** *contains exactly two dogs.*
Comparison between images	6	*There are* ***more*** *mammals in* ***the image on the right****.*

Table 10: Types of reasoning over the pair of images required in NLVR2, including the proportion of examples requiring each type and an example.

	Train	Dev	Test-P	Test-U
MAJORITY (assign True)	56.4/–	55.3/–	56.2/–	55.4/–
TEXT	58.4±0.6/–	56.6±0.5/–	57.2±0.6/–	56.2±0.4/–
IMAGE	56.8±1.3/–	55.4±0.1/–	56.1±0.3/–	55.3±0.3/–
CNN+RNN	58.9±0.2/–	56.6±0.3/–	58.0±0.3/–	56.3±0.6 /–
NMN	98.4±0.6/–	63.1±0.1/–	66.1±0.4/–	62.0±0.8/–
CNN-BIATT (Tan and Bansal, 2018)	–	66.9/–	69.7/–	**66.1/28.9**
W-MEMNN (Pavez et al., 2018)	–	65.6/–	65.8/–	–
CMM (Yao et al., 2018)	–	68.0/–	**69.9**/–	–
N2NMN (Hu et al., 2017):				
N2NMN-CLONING	95.6±1.3/79.9±4.7	57.9±1.1/9.7±0.8	–	–
N2NMN-TUNING	97.5±0.4/92.7±2.6	58.7±1.4/11.6±0.8	–	–
N2NMN-RL	95.4±2.4/81.2±10.6	65.3±0.4/16.2±1.5	69.1/20.7	66.0/17.7
FiLM (Perez et al., 2018)	95.5±0.4/84.6±2.7	60.1±1.2/14.6±1.3	62.2/18.4	61.2/18.1
MAC (Hudson and Manning, 2018)	64.2±4.7/12.6±0.2	55.4±0.5/7.4±0.6	57.6/11.7	54.3/8.6
HUMAN (approximation)	–	94.6±3.5/–	95.4±3.4/–	94.9±3.6/–

Table 11: Performance (accuracy/consistency) on NLVR.

series $[8192, 4096, 2048, 1024, 512, 256, 128, 64, 32, 16, 2]$.

E.1 Single Modality

TEXT The caption's representation is computed using an RNN encoder. We use 300-dimensional GloVe vectors trained on Common Crawl as word embeddings (Pennington et al., 2014). We encode the caption using a single-layer long short-term memory (LSTM, Hochreiter and Schmidhuber, 1997) RNN of size 4096. The hidden states of the caption are averaged and processed with the MLP described above to predict the truth value.

IMAGE The image pair's representation is computed by extracting features from a pre-trained model. We resize and pad each image with whitespace to a size of 530×416 pixels, which is the size of the image displayed to the workers during sentence-writing. Each padded image is resized to 224×224 and passed through a ResNet-152 pre-trained model (He et al., 2016). The features from the final layer before classification are extracted for each image and concatenated. This representation is processed with the MLP described above to predict a truth value.

E.2 Image and Text Baselines

CNN+RNN The caption and image pair are encoded as described in Appendix E.1, then concatenated and passed through the MLP described above to predict a truth value.

MAXENT We use n-grams where $2 \leq n \leq 6$. We train a maximum entropy classifier with Megam.[5]

E.3 Module Networks

End-to-End Neural Module Networks We use the publicly available implementation.[6] The model parameters used for NLVR2 are the same as those used for the original experiments on VQA. We use GloVe vectors of size 300 to embed words (Pennington et al., 2014). The model parameters used for NLVR are the same as those used for the original N2NMN experiments on CLEVR. This includes learning word embeddings from scratch and embedding images using the *pool5* layer of VGG-16 trained on ImageNet (Simonyan and Zisserman, 2014; Hu et al., 2017). The two paired images are resized and padded

[5] https://www.umiacs.umd.edu/~hal/megam

[6] https://github.com/ronghanghu/n2nmn

with white space to size 530×416, then concatenated horizontally and resized to a single image of 448×448 pixels. The resulting image is embedded using the *res5c* layer of ResNet-152 trained on ImageNet (He et al., 2016; Hu et al., 2017).

FiLM We use the publicly available implementation.[7] For NLVR2, we first resize and pad both images with whitespace to images of size 530×416. The two images are concatenated horizontally and resized to a single image of 224×224 pixels. This image is passed through a ResNet-101 pretrained model and the features from the *conv4* layer are extracted (He et al., 2016; Perez et al., 2018). For NLVR, we resize images to 224×224 and use the raw pixels directly. The parameters of the models are the same as described in Perez et al. (2018)'s experiments on featurized images, except for the following: RNN hidden size of 1096, classifier projection dimension of size 256, final MLP hidden size of 512, and 28 feature maps. Using the original parameters did not result in significant differences in accuracy, while updates using our parameters were computed faster and the computation graph used less memory.

E.4 MAC

We use the implementation provided online.[8] For experiments on NLVR2, we adapt the image processing procedure. Both images are resized and padded with white space to images of size 530×416, then concatenated horizontally and resized to 224×224 pixels. We use the same image featurization approach used in Hudson and Manning (2018). For experiments on NLVR, we use the NLVR configuration provided in the repository.

E.5 Training

For the TEXT, IMAGE, and CNN+RNN methods on NLVR2, we perform updates using ADAM (Kingma and Ba, 2014) with a global learning rate of 0.0001. The weights and biases are initialized by sampling uniformly from $[-0.1, 0.1]$. All fully-connected and output layers use a learned bias term. For MAC, we use the same training setup as described in Hudson and Manning (2018), stopping early based on performance over the development set. For all other experiments, we use early stopping with patience, where patience is initially set to a constant and multiplied 1.01 at each epoch the validation accuracy improves over a global maximum. We use 5% of the training data as a validation set, which is not used to update model parameters. We choose a validation set such that unique sentences do not appear in both the validation and training sets. For FiLM and N2NMN, we set the initial patience to 30. For TEXT, IMAGE and CNN+RNN baselines, initial patience was set to 10. For MAXENT, we use at most 100 epochs.

[7] https://github.com/ethanjperez/film

[8] https://github.com/stanfordnlp/mac-network

F Additional Examples

Table 12 includes additional examples sampled from the training and development sets of NLVR2, as well as license information for each image. All images in this paper were sampled from websites known for hosting non-copyrighted images, for example Wikimedia.

G Lisence Information

Tables 13, 14, 15, and 15 detail license and attribution information for the images included in the main paper.

Image Pair	Sentence	Label
 *Kropsoq (CC BY-SA 3.0); subhv150 (Pixabay)*	*Two hot air balloons are predominantly red and have baskets for passengers.*	True
 *babasteve (CC BY 2.0); Yathin S Krishnappa (CC BY-SA 3.0)*	*All elephants have ivory tusks.*	False
 *NatashaG (Pixabay); Photoman (Pixabay)*	*There are entirely green apples among the fruit in the right image.*	True
 *Pedi68 (Pixabay); Andrea Schafthuizen (PDP)*	*The animal in the image on the right is standing on its hind legs.*	False
 *Ben & Katherine Sinclair (CC BY 2.0); Zhangzhugang (CC BY-SA 3.0)*	*One of the images contains one baby water buffalo.*	True
 *Pelikana (CC BY-SA 3.0); violetta (Pixabay)*	*The sled in the image on the left is unoccupied.*	False
 *Frans de Waal (CC BY 2.5); Adam Jones (CC BY-SA 3.0)*	*Each image shows two animals interacting, and one image shows a monkey grooming the animal next to it.*	True
 *Burtonpe (CC BY-SA 3.0); Ville de Montréal (CC BY-SA 3.0)*	*In 1 of the images, the oars are kicking up spray.*	False
 *Sarah and Jason (CC BY-SA 2.0); Sarah and Jason (CC BY-SA 2.0)*	*In one image, a person is standing in front of a roofed and screened cage area with three different colored parrots perched them.*	True
 *Petey21 (CC0); Santeri Viinamäki (CC BY-SA 4.0)*	*In one of the images there are at least two golf balls positioned near a hole with a golf flagpole inserted in it.*	False

Table 12: Additional examples from the training and development sets of NLVR2, including license information for each photograph beneath the pair and the label of the example.

Image	Attribution and License
	MemoryCatcher (CC0)
	Calabash13 (CC BY-SA 3.0)
	Charles Rondeau (CC0)
	Andale (CC0)

Table 13: License information for the images in Figure 1.

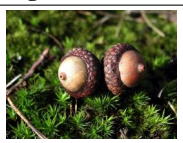
Image	Attribution and License
	Hagerty Ryan, USFWS (CC0)
	Charles Rondeau (CC0)
	Peter Griffin (CC0)
	Petr Kratochvil (CC0)
	George Hodan (CC0)
	Charles Rondeau (CC0)
	Andale (CC0)
	Maksym Pyrizhok (PDP)
	Sheila Brown (CC0)
	ulleo (CC0)

Table 14: License information for the images in Figure 2.

Image	Attribution and License
	JerryFriedman (CC0)
	Eric Kilby (CC BY-SA 2.0)
	Angie Garrett (CC BY 2.0)
	Ben HaTeva (CC BY-SA 2.5)
	Manfred Kopka (CC BY-SA 4.0)
	Aubrey Dale (CC BY-SA 2.0)
	Albert Bridge (CC BY-SA 2.0)
	Randwick (CC BY-SA 3.0)
	Alexas_Fotos (Pixabay)
	Alexas_Fotos (Pixabay)
	Ralph Daily (CC BY 2.0)
	hobbyknipse (Pixabay)

Table 15: License information for the images in Table 3.

Image	Attribution and License
	Nedih Limani (CC BY-SA 3.0)
	Jean-Pol GRANDMONT (CC BY-SA 3.0)
	Scott Robinson (CC BY 2.0)
	Tokumeigakarinoaoshima (CC0 1.0)
	CSIRO (CC BY 3.0)
	Dan90266 (CC BY-SA 2.0)
	Raimond Spekking (CC BY-SA 4.0)
	SamHolt6 (CC BY-SA 4.0)

Table 16: License information for the images in Table 2.

References

Manoj Acharya, Kushal Kafle, and Christopher Kanan. 2019. TallyQA: Answering complex counting questions. In *AAAI Conference on Artificial Intelligence*.

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.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian D. Reid, Stephen Gould, and Anton van den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683.

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.

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

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2019. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *IEEE Conference on Computer Vision and Pattern Recognition*.

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.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1080–1089.

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.

Nicholas FitzGerald, Yoav Artzi, and Luke Zettlemoyer. 2013. Learning distributions over logical forms for referring expression generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1914–1925.

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

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9.

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, 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.

Kushal Kafle and Christopher Kanan. 2017. An analysis of visual question answering algorithms. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1965–1973.

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.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*.

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

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. TVQA: Localized, compositional video question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379. Association for Computational Linguistics.

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*.

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

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.

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

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

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*.

Juan Pavez, Hector Alllende, and Hector Allende-Cid. 2018. Working memory networks: Augmenting memory networks with a relational reasoning module. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1000–1009.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

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.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

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.

Harm de Vries, Kurt Shuster, Dhruv Batra, Devi Parikh, Jason Weston, and Douwe Kiela. 2018. Talk the Walk: Navigating New York City through grounded dialogue. *CoRR*, abs/1807.03367.

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.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition*.

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.