Evaluating Visual Reasoning Through Grounded Language Understanding

Alane Suhr, Mike Lewis, James Yeh, Yoav Artzi

■ *Autonomous systems that understand natural language must reason about complex language and visual observations. Key to making progress toward such systems is the availability of benchmark data sets and tasks. We introduce the Cornell Natural Language Visual Reasoning (NLVR) corpus, which targets reasoning skills like counting, comparisons, and set theory. NLVR contains 92,244 examples of natural language statements paired with synthetic images and annotated with Boolean values for the simple task of determining whether the sentence is true or false about the image. While it presents a simple task, NLVR has been developed to challenge systems with diverse linguistic phenomena and complex reasoning. Linguistic analysis confirms that NLVR presents diversity and complexity beyond what is provided by contemporary benchmarks. Empirical evaluation of several methods further demonstrates the open challenges NLVR presents.*

Natural language provides an expressive framework to communicate about what we observe and do in the world. Resolving language meaning, then, requires considering not only sentences, but also the state of the world. Words refer to objects and attributes, phrases describe relations that define increasingly complex world states, and verbs map to actions that change one state to another. Building autonomous agents that understand natural language requires addressing this problem. Consider a household robot standing in front of the kitchen cabinet in figure 1, and asked to "take four of the larger plates from the middle shelf and put them on the table." This robot must identify the set of shelves, and then locate the middle one. To single out the set of plates to pick, it must not only count, but also compare sizes and shapes of objects. A key requirement for developing this type of reasoning is the availability of data and tasks to benchmark proposed approaches.

 ISSN 0738-4602

Take four of the larger plates from the middle shelf and put them on the table.

Figure 1. An Example Observation and Instruction Given to a Household Assistance Robot.

Indeed, significant efforts have been directed toward developing data sets that pose vision and language challenges (Antol et al. 2015; Chen et al. 2015). A key focus in existing resources has been diverse and realistic visual stimuli. For example, the Visual QA (VQA) data set (Antol et al. 2015) includes 265K COCO images (Lin et al. 2014), which contain dozens of object categories and over a million object instances. Questions were collected via crowdsourcing by asking workers to write questions given these images. While the collected questions are often challenging, answering them requires relatively rudimentary reasoning beyond the complex grounding problem. Understanding how well proposed approaches handle complex reasoning, including resolving numerical quantities, comparing sets, and reasoning about negated properties, remains an open challenge.

We address this challenge with the Cornell Natural Language Visual Reasoning (NLVR) data set (Suhr et al. 2017; Zhou, Suhr, and Artzi 2017). NLVR focuses on the problem of understanding complex, linguistically diverse natural language statements that require significant reasoning skills to understand. We design a simple task: given an image and a statement, the system must decide if the statement is true with regard to the image. Similar to VQA, and unlike caption generation, this binary classification task allows for straightforward evaluation. Figure 2 shows two examples from our data.

We use synthetic images to control the visual input during data collection. Each image shows an environment divided into three boxes. Each box contains various objects, either scattered about or stacked on one another. We use a small set of objects with few properties. This restriction enables us to simplify the recognition problem, and instead focus on reasoning about sets, counts, and spatial relations. The grouping into three sets is designed to support descriptions that contain set-theoretic language and numerical expressions.

The key challenge is collecting natural language descriptions that take advantage of the full complexity of the image, rather than focusing on simple properties, such as the existence of one object or another. The images support rich descriptions that include comparisons of sets, descriptions of spatial relations, counting of objects, and comparison of their properties. But how do we design a scalable process to collect such language?

Collecting the Data

We use crowdsourcing to collect descriptions from nonexperts. The key challenge is defining a task that will require the complexity of reasoning we aim to reflect. If we display a single image, workers will easily complete the task with sentences that contain simple references (for example, “there is a yellow triangle”). A key observation that underlies our process design is that discriminating between similar images is significantly harder and requires more complex reasoning. Furthermore, if instead of discriminating between images, the worker is asked to discriminate between sets of images, the task becomes more complex, and therefore requires the language to capture even finer distinctions.

These observations are at the foundation of a simple, yet surprisingly effective, data collection process. We generate four images to collect a description. We first generate two images separately by randomly sampling the number of objects and their properties. For each of the two images, we generate an additional image by shuffling the objects across the image. This gives us two pairs. The first pair includes the ini-

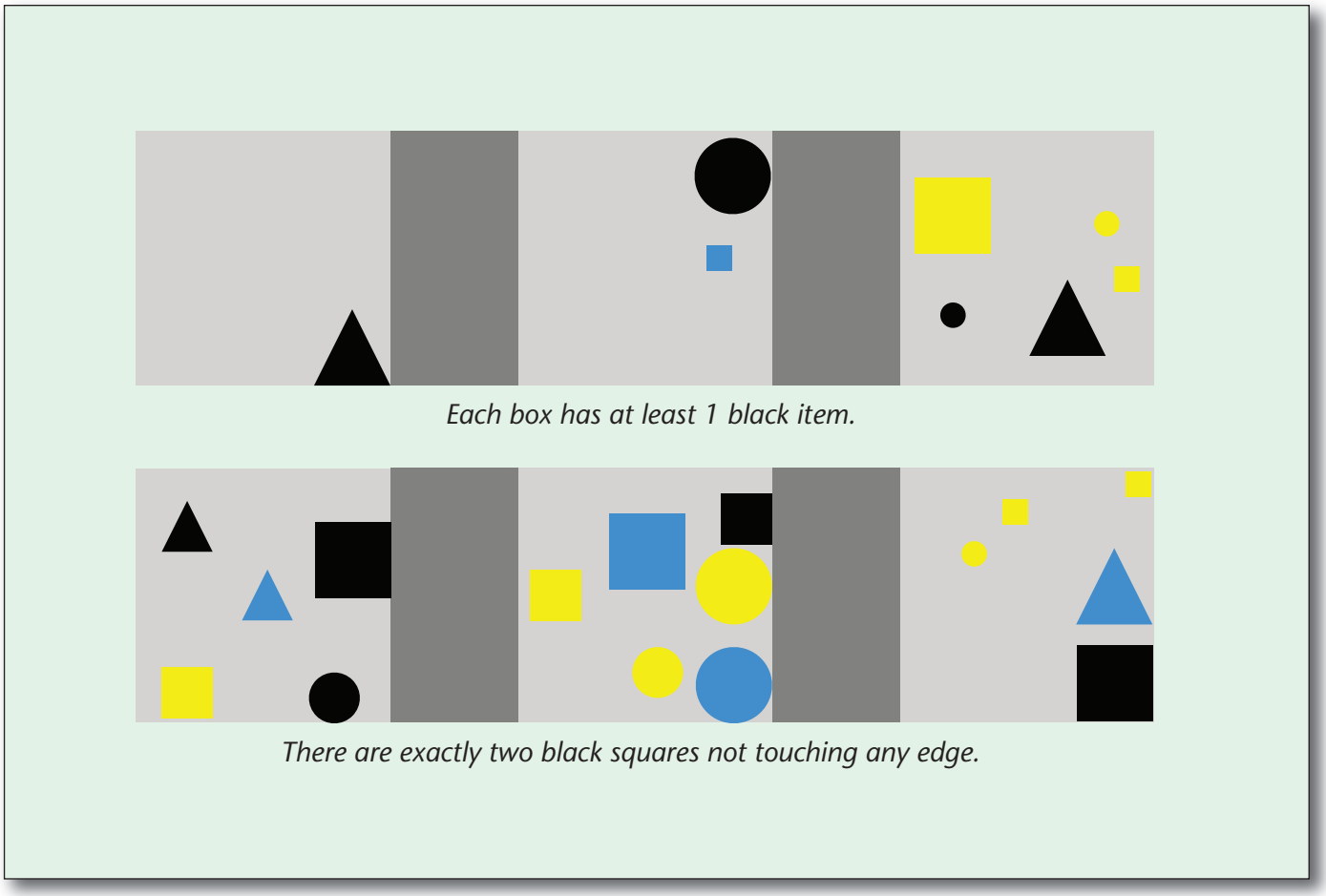


Figure 2. Example Sentences and Images from NLVR.

Each image includes three boxes with different object types. The truth value of the top sentence is true, while the bottom is false.

tial randomly generated images. The second pair is made up of their shuffled versions. We then ask the worker to write a sentence that is true for each of the images in the first pair, but false for each of the images in the second pair. To complete this task, workers must identify similarities between the first pair of images that do not hold for the second pair. The complexity of the task encourages language that expresses complex reasoning. Generating the second pair of images by shuffling the objects in the first pair prevents sentences that simply state the presence of a specific object. Our task encourages workers to write linguistically diverse and complex sentences by juxtaposing images that are similar to one another, yet contain minor differences.

We also asked the workers to follow two additional constraints. First, the sentence should not contain references to the image labels themselves. Second, the sentence should not refer to the horizontal order of the boxes. Treating the image as a set of three unordered boxes encourages set-theoretic descriptions. In addition to improving the language collected, these two constraints also allow us to generate a large number of examples. We can divide the results of each task into four independent image-sentence pairs, and then generate six images for each labeled sentence-image pair by permuting the boxes while maintaining the description's truth value. Figure 3 shows the prompt that was presented to the user.

This process already provides high-quality data. We quantify it by measuring agreement among annotators asked to solve the task given a sentence-image pair. We present annotators with an image and a sentence, and ask them to judge whether the sentence is true or false about the image. We also allow workers to mark examples as invalid. To validate the constraints, we randomly permute the boxes in the image before displaying it to the user. To compute agreement, we collect five judgments for examples in the development and test sets, and compute Krippendorf's α and Fleiss' κ (Cocos et al. 2015), two common agreement statistics. Our process yields $\alpha = 0.768$ and $\kappa = 0.709$, indicating substantial agreement (Landis and Koch 1977). We further increase the data quality by pruning examples that were marked as invalid and, for development and test

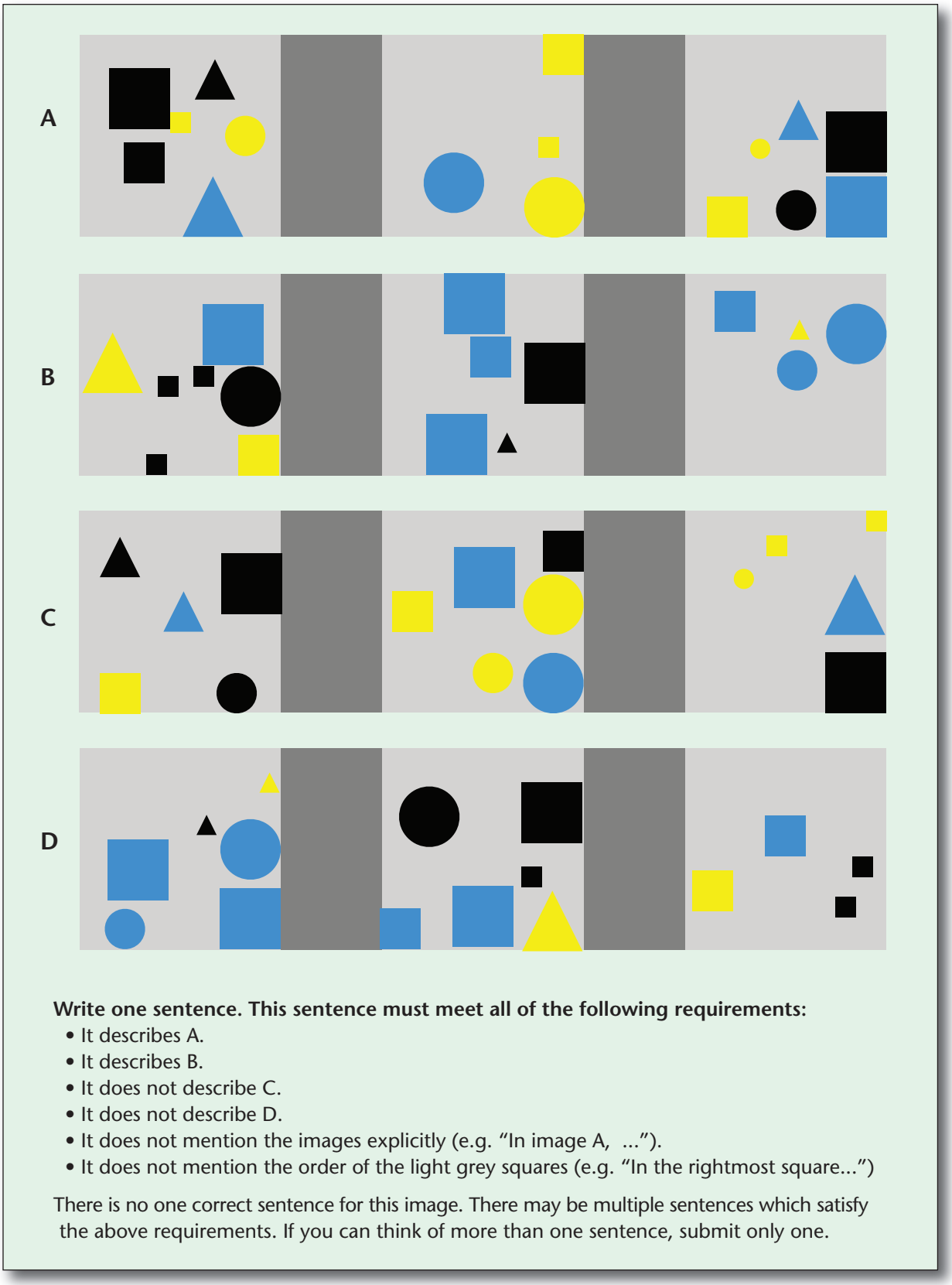


Figure 3. The NLVR Sentence Writing Prompt.

The bottom sentence in figure 2 was generated from this prompt. (Suhr et al. 2017, ©2017 ACL, reprinted with permission.)

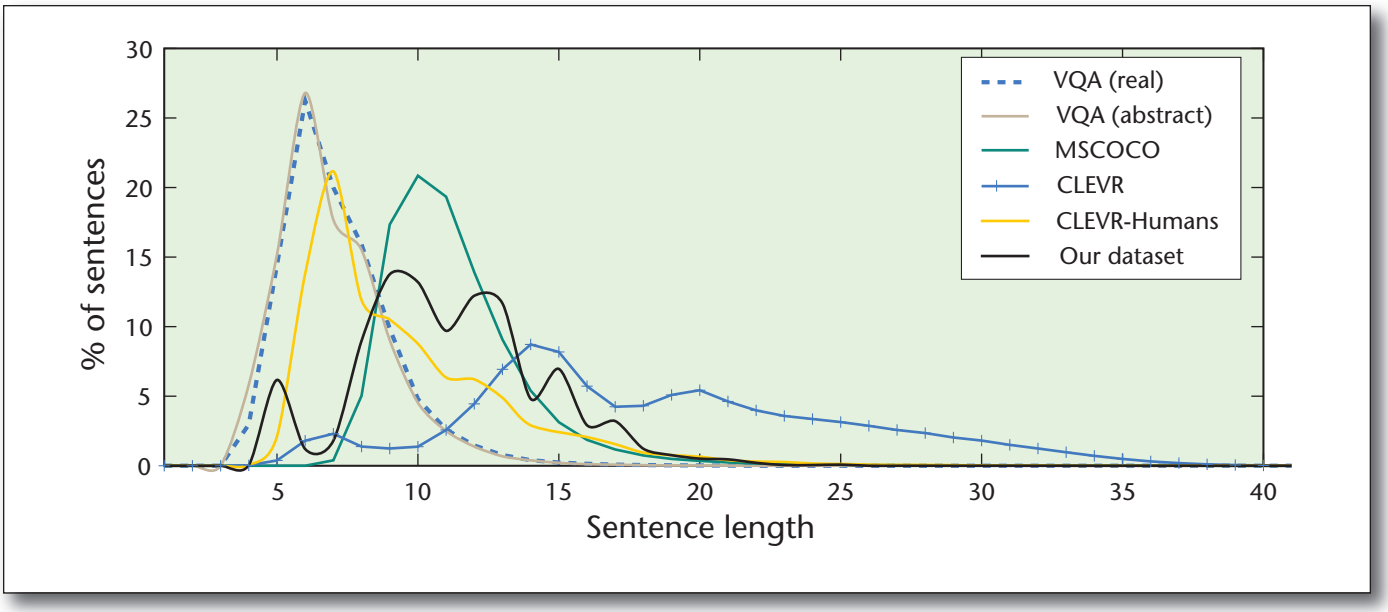


Figure 4. Distribution of Sentence Lengths.

examples with multiple labels, where disagreement is high. In practice, the pruning process removes only 3.3 percent of the original data. However, it increases agreement to $\alpha = 0.831$ and $\kappa = 0.808$.

We use the crowdsourcing platform Upwork.[1] We collected 92,244 image-sentence pairs labeled with whether the sentence is true about the image. The data contains 3,962 unique sentences. We split the data into four sets: a training set containing 80.7 percent of examples, a development set containing 6.4 percent of examples, and two test sets each containing 6.4 percent of examples. We keep one test set as unreleased, and use it to maintain a leaderboard.[2] We invite everyone working on the data to submit their models for evaluation on the unreleased set. Performance on the unreleased set is listed on the public leaderboard.

What Kind of Data Do We Get?

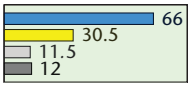
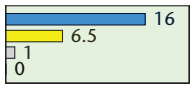
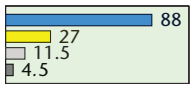
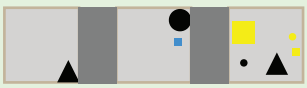
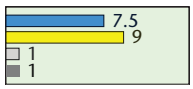
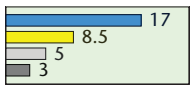
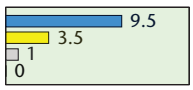
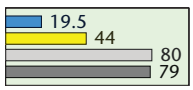
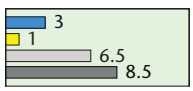
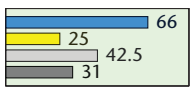
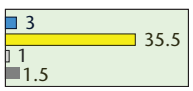
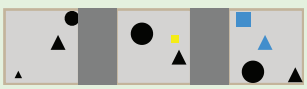
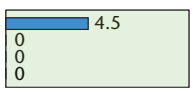
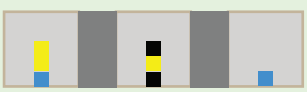
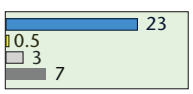
Our goal with NLVR is representation of linguistic diversity and complex reasoning. In an attempt to gain insight into the data we collected, we perform linguistic analysis of the data and compare our findings with several existing, related corpora. Our comparison focuses on VQA (Antol et al. 2015), which contains natural language questions about real photographs and synthetic abstract images; Microsoft COCO Captions (Chen et al. 2015), which contains natural language captions of photographs; and CLEVR (Johnson et al., CLEVR: A Diagnostic Dataset, 2017; Johnson et al., Inferring and Executing Programs, 2017), which contains both synthetic (CLEVR) and, more recently, human-written (CLEVR-Humans) questions about synthetic images.

We observe that sentence length in NLVR follows a similar distribution to the captions of Microsoft COCO Captions (figure 4). Longer sentences are often more challenging to understand, and display more compositionality. NLVR sentences are on average longer than those in VQA and CLEVR-Humans, but shorter than the synthetic sentences of CLEVR. We suspect the synthetic CLEVR sentences are longer due to the setup of the generation process.

We also study the presence of various linguistic phenomena in NLVR and the related corpora. This analysis is key to understanding the linguistic diversity and the type of reasoning required to solve the task. For example, a corpus with no reference to numbers or comparatives is likely not to require much cardinal reasoning. We choose 12 linguistic features directly related to our original goals, including counting, references to sets, and spatial relations. Table 1 lists the features we study, along with examples and their frequency in NLVR, VQA, and CLEVR-Humans. We analyze 200 examples in each corpus. We find that NLVR is remarkably diverse when compared to existing vision and language resources. For 10 out of the 12 categories, it shows higher representation than VQA. Even when compared to CLEVR-Humans, which was designed with a similar goal of benchmarking visual reasoning, NLVR contains more occurrences for 9 of the 12 features.

Biases, Baselines, and Challenges

The linguistic complexity of NLVR indicates that a variety of skills are required to solve the task. But how challenging is NLVR to existing methods? And what can we learn about the corpus from the per-

Feature	Sentence	Image	Comparison
Hard Cardinality	*There are **exactly two towers** with a yellow block at the top*		
Soft Cardinality	*There is **at least 1** yellow item in each box*		
Existential Quantifiers	***There are** two black blocks as the top of a tower*		
Universal Quantifiers	***Each box** has at least 1 black item*		
Coordination	*There is a box that has four items **and** the three are touching the side*		
Negation	*There are at least three yellow triangles **not** touching any edge*		
Presupposition	***The tower with two blocks** has a yellow block at the top*		
Coreference	*There is a tower with exactly three blocks, and **it** has a yellow block and two blue blocks*		
Spatial Relations	*There is at least one black object **above** a blue object*		
Comparative	*There are only two towers which has **the same base color***		
Coordination Ambiguity	*There is a box with exactly two blue items **and** at least two black items*		
Preposition Ambiguity	*There is a tower with a yellow block below a yellow block **at the top***		

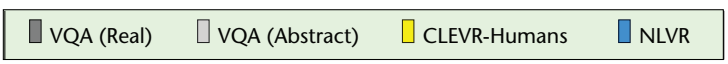


Table 1: Analysis of 12 Linguistic Features in NLVR, VQA (Real and Abstract Images), and CLEVR-Humans.

For each feature, we include an example containing a sentence and an image from NLVR. The bold text marks the occurrence of the feature in the example sentence. We also include the frequency of the different features in the analyzed corpora as computed by analyzing 200 sentences from each corpus.

formance of current approaches?

First, though, we study the corpus to identify latent biases. A common bias in vision and language corpora is an ability to solve the task with only one of the modalities. For example, this bias was recently identified in VQA, where several models achieved

high performance while ignoring the input image (Zhou et al. 2015; Jabri, Joulin, and van der Maaten 2016; Agrawal, Batra, and Parikh 2016; Kafle and Kanan 2017).[3] Does NLVR suffer from such a bias? NLVR is relatively balanced. Simply guessing true gives an accuracy of 55.4 percent on the unreleased test set. Using only one of the modalities provides similar results to this majority baseline. Encoding the image only with a convolutional neural network (CNN) to predict the truth values results in an accuracy of 55.3 percent. Similarly, encoding the text only with a recurrent neural network (RNN) to predict the truth value results in 56.2 percent accuracy. These results indicate that both the text and the image are necessary to solve the task.

A simple baseline that uses both text and image, however, provides disappointing results. We showed this by concatenating the outputs of the CNN and RNN models to predict the truth value. This model achieves 56.3 percent accuracy on the unreleased test set. In contrast, the neural module networks (NMN) approach (Andreas et al. 2016) achieves 62.0 percent. While performance is still low, this first success at outperforming the majority baseline is quite interesting. NMNs explicitly model compositionality. Different neural networks are composed together according to the structure of the sentence to process the image and generate the final prediction. The higher performance of this model indicates that understanding highly compositional language is necessary for solving the task.

An interesting property of NLVR is the availability of structured representations of the images. When generating images, we first generate a structured representation, which is then rendered to create the image. This representation contains the complete information about an image, including the items contained in the boxes, their properties, and exact positions. This representation can be considered as a small spatial database describing the environment, and enables experiments that do not require solving the vision problem.

Experimenting with the structured representation confirms an important property of the problem: counting is a necessary skill for solving NLVR. We use the sentence and structured representation to compute features and train a maximum entropy classifier. The classifier achieves an accuracy of 67.8 percent on the unreleased test set. Ablating all the features that consider counts reduces performance on the development set from 68.0 percent to 57.5 percent. This result clearly indicates the importance of counting in solving the task.

Treating the structured representation as a small database creates an interesting opportunity for semantic parsing techniques, where sentences are mapped to symbolic representations (as shown, for example, by Zelle and Mooney [1993], Zettlemoyer and Collins [2005], Zettlemoyer and Collins [2007], Clarke et al. [2010], Artzi and Zettlemoyer [2011], and Artzi and Zettlemoyer [2013]). Semantic parsing directly models the compositionality that is core to NLVR. As expected, semantic parsing can perform quite well on NLVR. This was shown recently by Goldman et al. (2017). Their approach maps the sentence into a small program, which is then executed against the structured representation to return a truth value. Their approach achieves 82.5 percent accuracy on the unreleased set. These results show that using a symbolic compositional representation improves performance on the task.

Discussion

Developing systems that rely on robust language and vision understanding requires data sets and tasks to evaluate their performance. The goal of NLVR is to present a challenging benchmark with linguistically diverse language that requires complex reasoning skills. Key to building NLVR is a carefully designed data collection process. The goal of the process is to challenge annotators to write sentences distinguishing several images. A study of the corpus shows it is more linguistically diverse compared to contemporary corpora. Our empirical analysis illustrates key challenges that must be addressed to solve NLVR, including counting and compositionality. While NLVR presents open challenges to the research community, its complexity is relatively scoped by our use of synthetically generated images with a limited number of shapes and properties. While we hope that NLVR will facilitate developing models that can better reason about vision and language, real-world applications require studying realistic visual inputs. An important direction we are currently pursuing is collecting a corpus that includes real images while preserving the complexity and diversity NLVR demonstrates. NLVR and leaderboards for both the image and structured representations are available.[4]

Notes

1. www.upwork.com.

2. The leaderboard is at lic.nlp.cornell.edu/nlvr.

3. Partially to address this bias, a new version of VQA was recently released (Goyal et al. 2017).

4. lic.nlp.cornell.edu/nlvr.

References

Agrawal, A.; Batra, D.; and Parikh, D. 2016. Analyzing the Behavior of Visual Question Answering Models. arXiv Preprint. arXiv:1606.07356v2 [cs.CL]. Ithaca, NY: Cornell University Library.

Andreas, J.; Rohrbach, M.; Darrell, T.; and Klein, D. 2016. Neural Module Networks. In *Proceedings of the 2016 IEEE Conference on Conference on Computer Vision and Pattern Recognition*. Los Alamitos, CA: IEEE Computer Society.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: Visual Question Answering. *International Journal of Computer Vision* 123(1): 4–31.

Artzi, Y., and Zettlemoyer, L. 2011. Bootstrapping Semantic Parsers from Conversations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing,* 421–432. Stroudsburg, PA: Association for Computational Linguistics.

Artzi, Y., and Zettlemoyer, L. 2013. Weakly Supervised Learning of Semantic Parsers for Mapping Instructions to Actions. *Transactions of the Association of Computational Linguistics* 1(1): 49–62.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. arXiv Preprint. arXiv:1504.00325 [cs.CV]. Ithaca, NY: Cornell University Library.

Clarke, J.; Goldwasser, D.; Chang, M.-W.; and Roth, D. 2010. Driving Semantic Parsing from the World's Response. In *Proceedings of the 14th Conference on Computational Natural Language Learning*. Stroudsburg, PA: Association for Computational Linguistics.

Cocos, A.; Masino, A.; Qian, T.; Pavlick, E.; and Callison-Burch, C. 2015. Effectively Crowdsourcing Radiology Report Annotations. In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*, 109–114. Stroudsburg, PA: Association for Computational Linguistics.

Goldman, O.; Latcinnik, V.; Naveh, U.; Globerson, A.; and Berant, J. 2017. Weakly-Supervised Semantic Parsing with Abstract Examples. arXiv Preprint. arXiv:1711.05240 [cs.CL]. Ithaca, NY: Cornell University Library.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Proceedings of the 2017 IEEE Conference on Conference on Computer Vision and Pattern Recognition*. Los Alamitos, CA: IEEE Computer Society.

Jabri, A.; Joulin, A.; and van der Maaten, L. 2016. Revisiting Visual Question Answering Baselines. In *Computer Vision – ECCV 2016: 14th European Conference*. Lecture Notes in Computer Science 9905. Berlin: Springer.

Johnson, J.; Hariharan, B.; van der Maaten, L.; Fei Li, F.; Zitnick, C. L.; and Girshick, R. B. 2017. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*. Los Alamitos, CA: IEEE Computer Society.

Johnson, J.; Hariharan, B.; van der Maaten, L.; Hoffman, J.; Fei Li, F.; Zitnick, C. L.; and Girshick, R. B. 2017. Inferring and Executing Programs for Visual Reasoning. In *Proceedings of the IEEE International Conference on Computer Vision*. Los Alamitos, CA: IEEE Computer Society.

Kafle, K., and Kanan, C. 2017. Visual Question Answering: Datasets, Algorithms, and Future Challenges. *Computer Vision and Image Understanding* 163: 3–20.

Landis, J. R., and Koch, G. G. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33(1):159–74.

Lin, T.-Y.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014: 13th European Conference*. Lecture Notes in Computer Science 8689. Berlin: Springer.

Suhr, A.; Lewis, M.; Yeh, J.; and Artzi, Y. 2017. A Corpus of Natural Language for Visual Reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics 2*, 217–223. Stroudsburg, PA: Association for Computational Linguistics.

Zelle, J. M., and Mooney, R. J. 1993. Learning Semantic Grammars with Constructive Inductive Logic Programming. In *Proceedings of the 11th National Conference on Artificial Intelligence*, 817–822. Menlo Park, CA: AAAI Press.

Zettlemoyer, L. S., and Collins, M. 2005. Learning to Map Sentences to Logical Form: Structured Classification with Probabilistic Categorial Grammars. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence*, 658–666. Seattle, WA: AUAI Press.

Zettlemoyer, L. S., and Collins, M. 2007. Online Learning of Relaxed CCG Grammars for Parsing to Logical Form. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 678–687. Stroudsburg, PA: Association for Computational Linguistics.

Zhou, S.; Suhr, A.; and Artzi, Y. 2017. Visual Reasoning with Natural Language. Paper presented at the AAAI 2017 Fall Symposium on Natural Communication for Human-Robot Collaboration. Arlington, VA, Nov. 9–11.

Zhou, B.; Tian, Y.; Sukhbaatar, S.; Szlam, A.; and Fergus, R. 2015. Simple Baseline for Visual Question Answering. arXiv Preprint. arXiv:1512.02167 [cs.CV]. Ithaca, NY: Cornell University Library.

Alane Suhr is a PhD student in the Department of Computer Science at Cornell Tech, Cornell University, focusing on building agents that understand natural language grounded in complex interactions. She is the recipient of an AI2 Key Scientific Challenges Award, a Microsoft Research Women's Fellowship, a Best Paper award at ACL 2017, and an Outstanding Paper award at NAACL 2018. Suhr received a bachelor's degree in computer science and engineering from Ohio State University in 2016.

Mike Lewis is a scientist at Facebook AI Research, working on connecting language and reasoning. Previously, he was a postdoc at the University of Washington, developing search algorithms for neural structured prediction. Lewis has a PhD from the University of Edinburgh on combining symbolic and distributed representations of meaning.

James Yeh is a software engineering at Evidation Health, working on enabling people to participate in better health outcomes. Yeh received his master's in operations research and information engineering at Cornell Tech, where he contributed to NLP research under the supervision of Yoav Artzi and pursued his interests in AI and using data to enhance decision making. He received his bachelor's degree in applied science under the Department of Systems Design Engineering at the University of Waterloo, where he focused on the study of intelligent systems.

Yoav Artzi is an assistant professor in the Department of Computer Science at Cornell Tech, Cornell University. His research focuses on learning expressive models for natural language understanding, most recently in situated interactive scenarios. He received an NSF CAREER award, Best Paper awards in EMNLP 2015 and ACL 2017, and a Google faculty award. Artzi holds a BSc summa cum laude from Tel Aviv University and a PhD from the University of Washington.