Measuring Machine Intelligence Through Visual Question Answering

C. Lawrence Zitnick, Aishwarya Agrawal, Stanislaw Antol, Margaret Mitchell, Dhruv Batra, Devi Parikh

■ *As machines have become more intelligent, there has been a renewed interest in methods for measuring their intelligence. A common approach is to propose tasks for which a human excels, but one that machines find difficult. However, an ideal task should also be easy to evaluate and not be easily gameable. We begin with a case study exploring the recently popular task of image captioning and its limitations as a task for measuring machine intelligence. An alternative and more promising task is visual question answering, which tests a machine's ability to reason about language and vision. We describe a data set, unprecedented in size and created for the task, that contains more than 760,000 human-generated questions about images. Using around 10 million human-generated answers, researchers can easily evaluate the machines.*

Humans have an amazing ability to both understand and reason about our world through a variety of senses or modalities. A sentence such as "Mary quickly ran away from the growling bear" conjures both vivid visual and auditory interpretations. We picture Mary running in the opposite direction of a ferocious bear with the sound of the bear being enough to frighten anyone. While interpreting a sentence such as this is effortless to a human, designing intelligent machines with the same deep understanding is anything but. How would a machine know Mary is frightened? What is likely to happen to Mary if she doesn't run? Even simple implications of the sentence, such as "Mary is likely outside" may be nontrivial to deduce.

How can we determine whether a machine has achieved the same deep understanding of our world as a human? In our example sentence above, a human's understanding is rooted in multiple modalities. Humans can visualize a scene depicting Mary running, they can imagine the sound of the bear, and even how the bear's fur might feel when touched. Conversely, if shown a picture or even an auditory recording of a woman running from a bear, a human may similarly describe the scene. Perhaps machine intelligence could be tested in a similar manner? Can a machine use natural language to describe a picture similar to a human? Similarly, could a machine generate a scene given a written description? In fact these tasks have been a goal of artificial intelligence research since its inception. Marvin Minsky famously stated in 1966 (Crevier 1993) to one of his students, "Connect a television camera to a computer and get the machine

 ISSN 0738-4602

A man holding a beer bottle with two hands and looking at it.

A man in a white t-shirt looks at his beer bottle.

A man with black curly hair is looking at a beer.

A man holds a bottle of beer examining the label.

…

A guy holding a beer bottle.

A man holding a beer bottle.

A man holding a beer.

A man holds a bottle.

Man holding a beer.

Figure 1. Example Image Captions Written for an Image Sorted by Caption Length.

to describe what it sees." At the time, and even today, the full complexities of this task are still being discovered.

Image Captioning

Are tasks such as image captioning (Barnard and Forsyth 2001; Kulkarni et al. 2011; Mitchell et al. 2012; Farhadi et al. 2010; Hodosh, Young, and Hockenmaier 2013; Fang et al. 2015; Chen and Zitnick 2015; Donahue et al. 2015; Mao et al. 2015; Kiros, Salakhutdinov, and Zemel 2015; Karpathy and Fei-Fei 2015; Vinyals et al. 2015) promising candidates for testing artificial intelligence? These tasks have advantages, such as being easy to describe and being capable of capturing the imagination of the public (Markoff 2014). Unfortunately, tasks such as image captioning have proven problematic as actual tests of intelligence. Most notably, the evaluation of image captions may be as difficult as the image captioning task itself (Elliott and Keller 2014; Vedantam, Zitnick, and Parikh 2015; Hodosh, Young, and Hockenmaier 2013; Kulkarni et al. 2011; Mitchell et al. 2012). It has been observed that captions judged to be good by human observers may actually contain significant variance even though they describe the same image (Vedantam, Zitnick, and Parikh 2015). For instance see figures 1. Many people would judge the longer, more detailed captions as better. However, the details described by the captions vary significantly, for example, two hands, white T-shirt, black curly hair, label, and others. How can we evaluate a caption if there is no consensus on what should be contained in a *good* caption? However, for shorter, less detailed captions that are commonly written by humans, a rough consensus is achieved: "A man holding a beer bottle." This leads to the somewhat counterintuitive conclusion that captions humans like aren't necessarily humanlike.

The task of image captioning also suffers from another less obvious drawback. In many cases it might be too easy! Consider an example success from a recent paper on image captioning (Fang et al. 2015), figure 4. Upon first inspection this caption appears to have been generated from a deep understanding of the image. For instance, in figure 4 the machine must have detected a giraffe, grass, and a tree. It understood that the giraffe was standing, and the thing it was standing on was grass. It knows the tree and giraffe are *next to* each other, and others. Is this interpretation of the machine's depth of understanding correct? When judging the results of an AI system, it is important to analyze not only its output but also the data used for its training. The results in figure 4 were obtained by training on the Microsoft common objects in context (MS COCO) data set (Lin et al. 2014). This data set contains five independent captions written by humans for more than 120,000 images (Chen et al. 2015). If we examine the image in figure 4 and the images in the training data set we can make an interesting observation. For many testing images, there exist a significant number of semantically similar training images, figure 4 (right). If two images share enough semantic similarity, it is

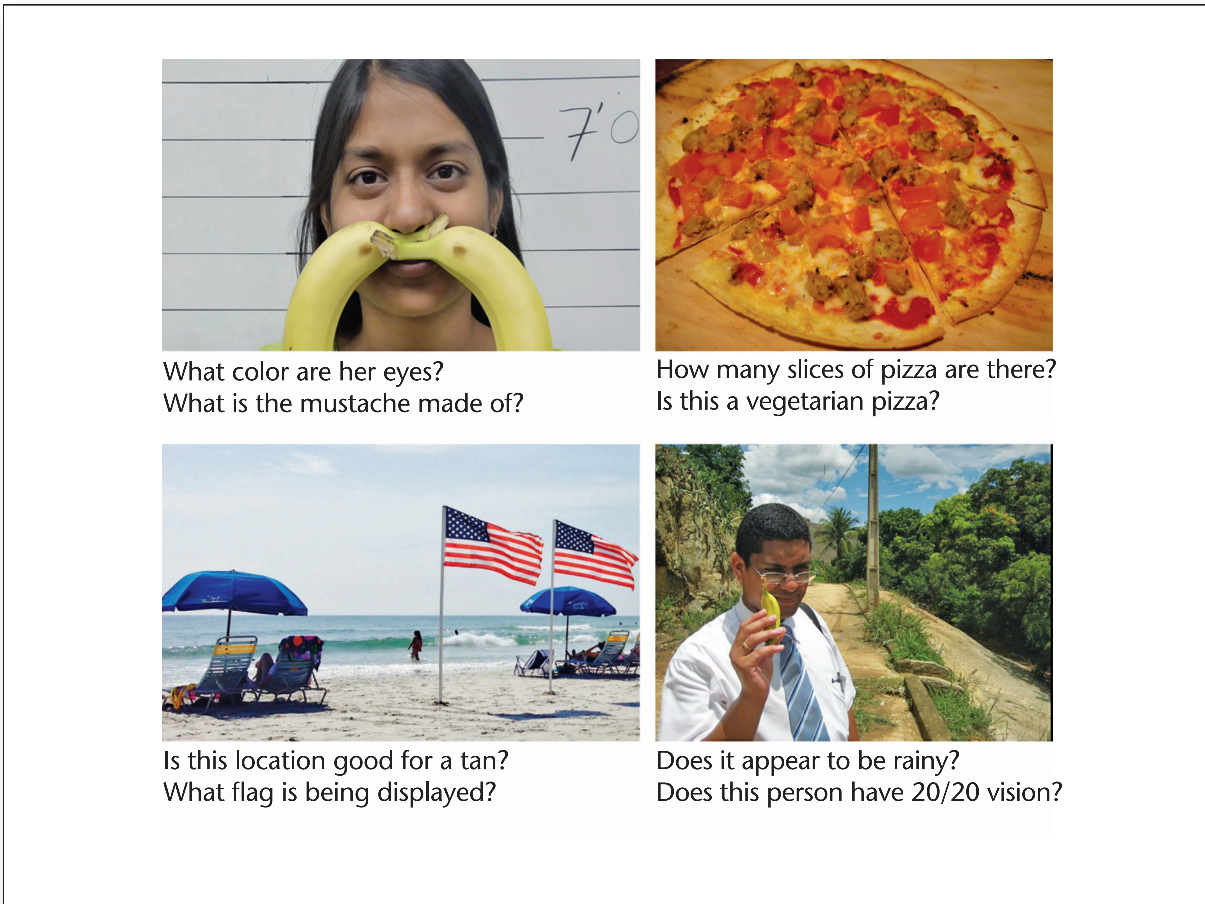

Figure 2. Example Images and Questions in the Visual Question-Answering Data Set. (visualqa.org).

possible a single caption could describe them both.

This observation leads to a surprisingly simple algorithm for generating captions (Devlin et al. 2015). Given a test image, collect a set of captions from images that are visually similar. From this set, select the caption with highest consensus (Vedantam, Zitnick, and Parikh 2015), that is, the caption most similar to the other captions in the set. In many cases the consensus caption is indeed a good caption. When judged by humans, 21.6 percent of these borrowed captions are judged to be equal to or better than those written by humans for the image specifically. Despite its simplicity, this approach is competitive with more advanced approaches that use recurrent neural networks (Chen and Zitnick 2015; Donahue et al. 2015; Mao et al. 2015; Kiros, Salakhutdinov, and Zemel 2015; Karpathy and Fei-Fei 2015; Vinyals et al. 2015) and other language models (Fang et al. 2015) that can achieve 27.3 percent when compared to human captions. Even methods using recurrent neural networks commonly produce captions that are identical to training captions even though they're not explicitly trained to do so. If captions are generated by borrowing them from other images, these algorithms are clearly not demonstrating a deep understanding of language, semantics, and their visual interpretation. In comparison, the odds of two humans repeating a sentence are quite rare.

One could make the case that the fault is not with the algorithms but in the data used for training. That is, the data set contains too many semantically similar images. However, even in randomly sampled images from the web, a photographer bias is found. Humans capture similar images to each other. Many of our tastes or preferences are conventional.

Visual Question Answering

As we demonstrated using the task of image captioning, determining a multimodal task for measuring a machine's intelligence is challenging. The task must be easy to evaluate, yet hard to solve. That is, its evaluation shouldn't be as hard as the task itself, and it must not be solvable using shortcuts or cheats. To solve these two problems we propose the task of visual question answering (VQA) (Antol et al. 2015; Geman et al. 2015; Malinowski and Fritz 2014; Tu et al. 2014; Bigham et al. 2010; Gao et al. 2015).

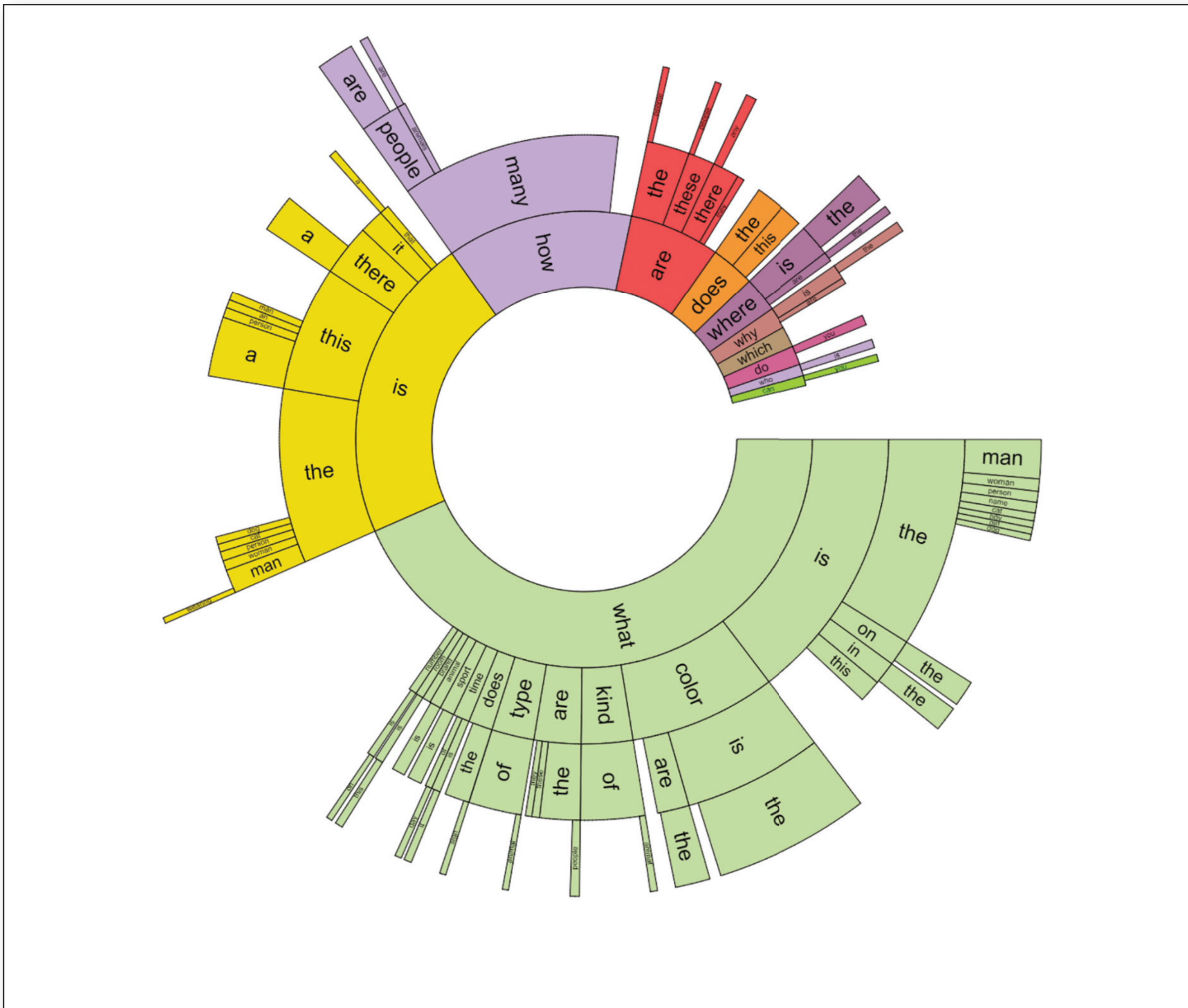

Figure 3. Distribution of Questions by Their First Four Words.

The ordering of the words starts toward the center and radiates outwards. The arc length is proportional to the number of questions containing the word. White areas indicate words with contributions too small to show.

The task of VQA requires a machine to answer a natural language question about an image as shown in figure 2. Unlike the captioning task, evaluating answers to questions is relatively easy. The simplest approach is to pose the questions with multiple choice answers, much like standardized tests administered to students. Since computers don't get tired of reading through long lists of answers, we can even increase the length of the answer list. Another more challenging option is to leave the answers open ended. Since most answers are single words such as *yes, blue,* or *two,* evaluating their correctness is straightforward.

Is the visual question-answering task challenging? The task is inherently multimodal, since it requires knowledge of language and vision. Its complexity is further increased by the fact that many questions require commonsense knowledge to answer. For instance, if you ask, "Does the man have 20/20 vision?" you need the commonsense knowledge that having 20/20 vision implies you don't wear glasses. Going one step further, one might be concerned that commonsense knowledge is all that's needed to answer the questions. For example if the question was "What color is the sheep?," our common sense would tell us the answer is *white*. We may test the sufficiency of commonsense knowledge by asking subjects to answer questions without seeing the accompanying image. In this case, human subjects did indeed perform poorly (33 percent correct), indicating that common sense may be necessary but is not sufficient. Similarly, we may ask subjects to answer the question given only a caption describing the image. In this case the humans performed better (57 percent correct), but still not as accurately as those able to view the image (78 percent correct). This helps indicate that the VQA task requires more

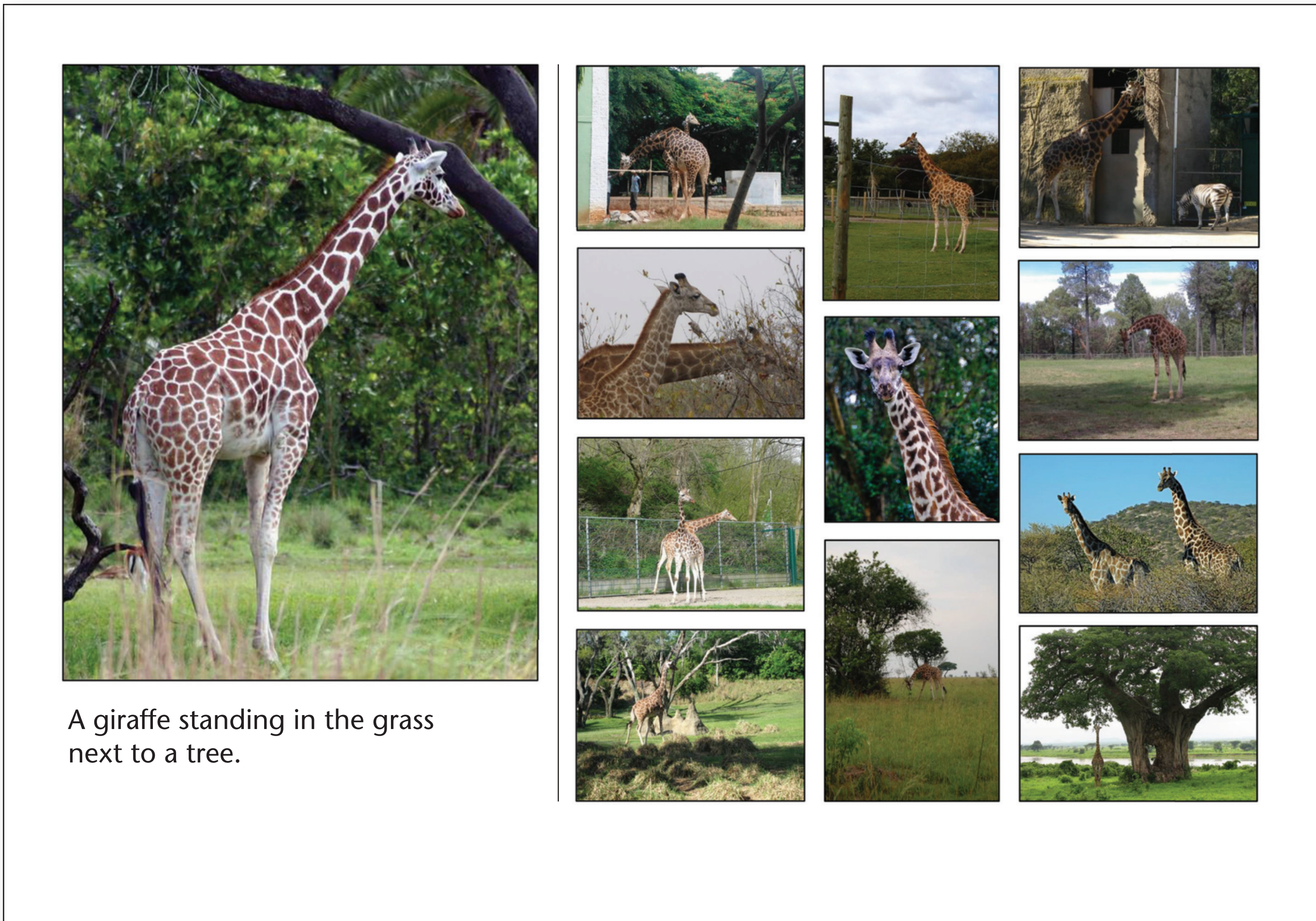

Figure 4. Example Image Caption and a Set of Semantically Similar Images.

Left: An image caption generated from Fang et al. (2015). *Right:* A set of semantically similar images in the MS COCO training data set for which the same caption could apply.

detailed information about an image than is typically provided in an image caption.

How do you gather diverse and interesting questions for 100,000s of images? Amazon's Mechanical Turk provides a powerful platform for crowdsourcing tasks, but the design and prompts of the experiments must be careful chosen. For instance, we ran trial experiments prompting the subjects to write questions that would be difficult for a toddler, alien, or smart robot to answer. Upon examination, we determined that questions written for a smart robot were most interesting given their increased diversity and difficulty. In comparison, the questions stumping a toddler were a bit too easy. We also gathered three questions per image and ensured diversity by displaying the previously written questions and stating, "Write a different question from those above that would stump a smart robot." In total over 760,000 questions were gathered.[1]

The diversity of questions supplied by the subjects on Amazon's Mechanical Turk is impressive. In figure 3, we show the distribution of words that begin the questions. The majority of questions begin with *What* and *Is,* but other questions include *How, Are, Does,* and others. Clearly no one type of question dominates. The answers to these questions have a varying diversity depending on the type of question. Since the answers may be ambiguous, for example, "What is the person looking at?" we collected 10 answers per question. As shown in figure 5, many question types are simply answered *yes* or *no*. Other question types such as those that start with "What is" have a greater variety of answers. An interesting comparison is to examine the distribution of answers when subjects were asked to answer the questions with and without looking at the image. As shown in Figure 5 (bottom), there is a strong bias to many questions when subjects do not see the image. For

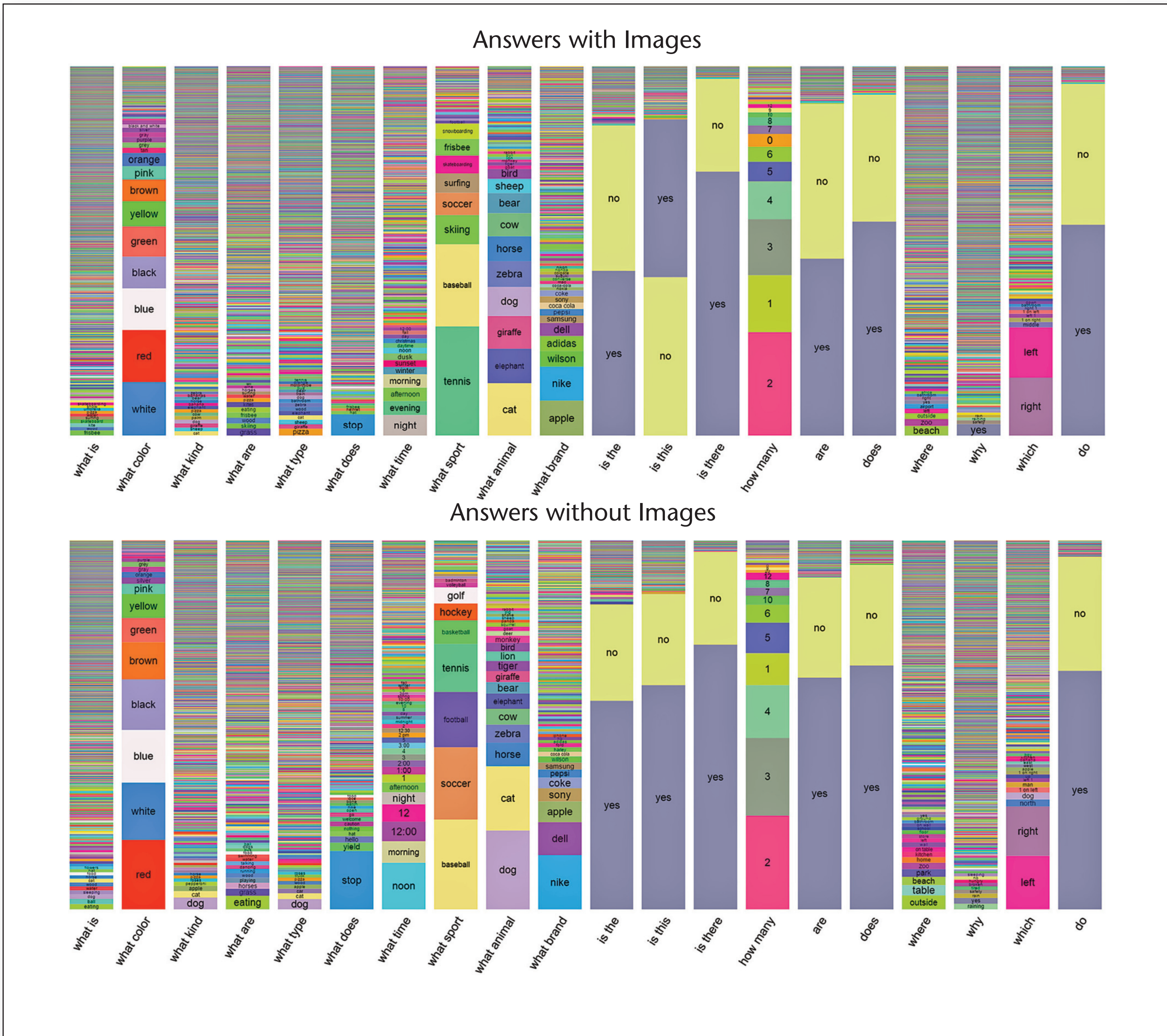

Figure 5. Distribution of Answers Per Question Type.

Top: When subjects provide answers when given the image. *Bottom:* When not given the image.

instance "What color" questions invoke red as an answer, or for questions that are answered by yes or no, yes is highly favored.

Finally it is important to measure the difficulty of the questions. Some questions such as "What color is the ball?" or "How many people are in the room?" may seem quite simple. In contrast, other questions such as "Does this person expect company?" or "What government document is needed to partake in this activity?" may require quite advanced reasoning to answer. Unfortunately, the difficultly of a question is in many cases ambiguous. The question's difficultly is as much dependent on the person or machine answering the question as the question itself. Each person or machine has different competencies.

In an attempt to gain insight into how challenging each question is to answer, we asked human subjects to guess how old a person would need to be to answer the question. It is unlikely most human subjects have adequate knowledge of human learning development to answer the question correctly. However, this does provide an effective proxy for question

3-4 (15.3%)	**5-8** (39.7%)	**9-12** (28.4%)	**13-17** (11.2%)	**18+** (5.5%)
Is that a bird in the sky?	How many pizzas are shown?	Where was this picture taken?	Is he likely to get mugged if he walked down a dark alleyway like this?	What type of architecture is this?
What color is the shoe?	What are the sheep eating?	What ceremony does the cake commemorate?	Is this a vegetarian meal?	Is this a Flemish bricklaying pattern?
How many zebras are there?	What color is his hair?	Are these boats too tall to fit under the bridge?	What type of beverage is in the glass?	How many calories are in this pizza?
Is there food on the table?	What sport is being played?	What is the name of the white shape under the batter?	Can you name the performer in the purple costume?	What government document is needed to partake in this activity?
Is this man wearing shoes?	Name one ingredient in the skillet.	Is this at the stadium?	Besides these humans, what other animals eat here?	What is the make and model of this vehicle?

Figure 6. Example Questions Judged to Be Answerable by Different Age Groups.

The percentage of questions falling into each age group is shown in parentheses.

difficulty. That is, questions judged to be answerable by a 3–4 year old are easier than those judged answerable by a teenager. Note, we make no claims that questions judged answerable by a 3–4 year old will actually be answered correctly by toddlers. This would require additional experiments performed by the appropriate age groups. Since the task is ambiguous, we collected 10 responses for each question. In Figure 6 we show several questions for which a majority of subjects picked the specified age range.

Surprisingly the perceived age needed to answer the questions is fairly well distributed across the different age ranges. As expected the questions that were judged answerable by an adult (18+) generally need specialized knowledge, where those answerable by a toddler (3–4) are more generic.

Abstract Scenes

The visual question-answering task requires a variety of skills. The machine must be able to understand the image, interpret the question, and reason about the answer. For many researchers exploring AI, they may not be interested in exploring the low-level tasks involved with perception and computer vision. Many of the questions may even be impossible to solve given the current capabilities of state-of-the-art computer vision algorithms. For instance the question "How many cellphones are in the image?" may not be answerable if the computer vision algorithms cannot accurately detect cellphones. In fact, even for state-of-the-art algorithms many objects are difficult to detect, especially small objects (Lin et al. 2014).

To enable multiple avenues for researching VQA, we introduce abstract scenes into the data set (Antol, Zitnick, and Parikh 2014; Zitnick and Parikh 2013; Zitnick, Parikh, and Vanderwende 2013; Zitnick, Vedantam, and Parikh 2015). Abstract scenes or cartoon images are created from sets of clip art, figure 7. The scenes are created by human subjects using a graphical user interface that allows them to arrange a wide variety of objects. For clip art depicting humans, their poses and expression may also be changed. Using the interface, a wide variety of scenes can be created including ordinary scenes, scary scenes, or funny scenes.

Since the type of clip art and its properties are exactly known, the problem of recognizing objects and their attributes is greatly simplified. This provides researchers an opportunity to study more directly the problems of question understanding and answering. Once computer vision algorithms catch up, perhaps some of the techniques developed for abstract scenes can be applied to real images. The abstract scenes may be useful for a variety of other tasks as well, such as learning commonsense knowledge (Zitnick, Parikh, and Vanderwende 2013; Antol, Zitnick, and Parikh 2014; Chen, Shrivastava, and Gupta 2013; Divvala, Farhadi, and Guestrin 2014; Vedantam et al. 2015).

Discussion

While visual question answering appears to be a promising approach to measuring machine intelligence for multimodal tasks, it may prove to have

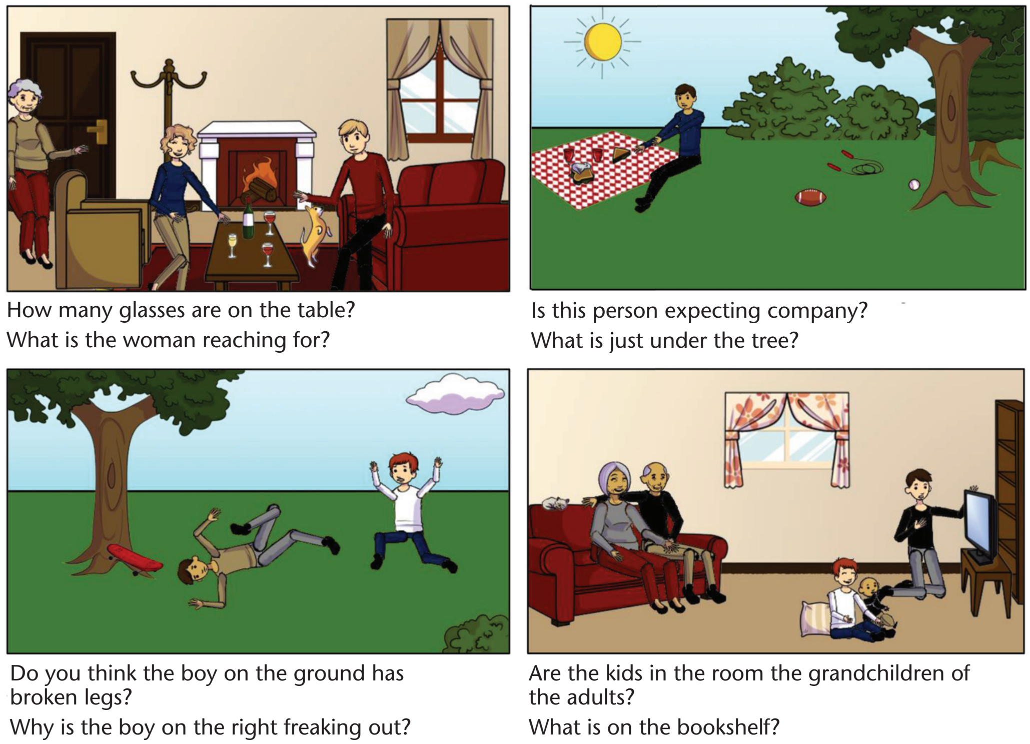

Figure 7. Example Abstract Scenes and Their Questions in the Visual Question-Answering Data Set.

visualqa.org.

unforeseen shortcomings. We've explored several baseline algorithms that perform poorly when compared to human performance. As the data set is explored, it is possible that solutions may be found that don't require true AI. However, using proper analysis we hope to update the data set continuously to reflect the current progress of the field. As certain question or image types become too easy to answer we can add new questions and images. Other modalities may also be explored such as audio and text-based stories (Fader, Zettlemoyer, and Etzioni 2013a, 2013b; Weston et al. 2014, Richardson, Burges, and Renshaw 2013).

In conclusion, we believe designing a multimodal challenge is essential for accelerating and measuring the progress of AI. Visual question answering offers one approach for designing such challenges that allows for easy evaluation while maintaining the difficultly of the task. As the field progresses our tasks and challenges should be continuously reevaluated to ensure they are of appropriate difficultly given the state of research. Importantly, these tasks should be designed to push the frontiers of AI research and help ensure their solutions lead us toward systems that are truly AI complete.

Notes

1. visualqa.org.

References

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: Visual Question Answering. Unpublished paper deposited in The Computing Research Repository (CoRR) 1505.00468. Association for Computing Machinery.

Antol, S.; Zitnick, C. L.; and Parikh, D. 2014. Zero-Shot Learning via Visual Abstraction. In *Computer Vision-ECCV 2014: Proceedings of the 13th European Conference,* Part IV. Lecture Notes in Computer Science Volume 8692. Berlin: Springer. dx.doi.org/10.1007/978-3-319-10593-2_27

Barnard, K., and Forsyth, D. 2001. Learning the Semantics of Words and Pictures. In *Proceedings of the IEEE International Conference on Computer Vision* (ICCV-01), 408–415. Los

Alamitos, CA: IEEE Computer Society. dx.doi.org/10.1109/iccv.2001.937654

Bigham, J.; Jayant, C.; Ji, H.; Little, G.; Miller, A.; Miller, R.; Miller, R.; Tatarowicz, A.; White, B.; White, S.; and Yeh, T. 2010. VizWiz: Nearly Real-Time Answers to Visual Questions. In *Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology*. New York: Association for Computing Machinery.

Chen, X., and Zitnick, C. L. 2015. Mind's Eye: A Recurrent Visual Representation for Image Caption Generation. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: Institute for Electrical and Electronics Engineers. dx.doi.org/10.1109/CVPR.2015.7298856

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. Unpublished paper deposited in The Computing Research Repository (CoRR) 1504.00325. Association for Computing Machinery.

Chen, X.; Shrivastava, A.; and Gupta, A. 2013. NEIL: Extracting Visual Knowledge from Web Data. In *Proceedings of the IEEE International Conference on Computer Vision,* ICCV 2013. Piscataway, NJ: Institute for Electrical and Electronics Engineers. dx.doi.org/10.1109/iccv.2013.178

Crevier, D. 1993. *AI: The Tumultuous History of the Search for Artificial Intelligence*. New York: Basic Books, Inc.

Devlin, J.; Gupta, S; Girshick, R.; Mitchell, M.; and Zitnick, C. L. 2015. Exploring Nearest Neighbor Approaches for Image Captioning. Unpublished paper deposited in The Computing Research Repository (CoRR) 1505.04467. Association for Computing Machinery.

Divvala, S.; Farhadi, A.; and Guestrin, C. 2014. Learning Everything About Anything: Webly-Supervised Visual Concept Learning. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: Institute for Electrical and Electronics Engineers. dx.doi.org/10.1109/CVPR.2014.412

Donahue, J.; Hendricks, L. A.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; and Darrell, T. 2015. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: Institute for Electrical and Electronics Engineers. dx.doi.org/10.1109/CVPR.2015.7298878

Elliott, D., and Keller, F. 2014. Comparing Automatic Evaluation Measures for Image Description. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Stroudsburg PA: Association for Computational Linguistics. x.doi.org/10.3115/v1/p14-2074

Fader, A.; Zettlemoyer, L.; and Etzioni, O. 2013a. Open Question Answering over Curated and Extracted Knowledge Bases. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: Assocation for Computing Machinery.

Fader, A.; Zettlemoyer, L.; and Etzioni, O. 2013b. Paraphrase-Driven Learning for Open Question Answering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics.Engineers.

Farhadi, A.; Hejrati, M.; Sadeghi, M. A.; Young, P.; Rashtchian, C.; Hockenmaier, J.; and Forsyth, D. 2010. Every Picture Tells a Story: Generating Sentences from Images. In *Computer Vision–ECCV 2010, Proceedings of the 11th European Conference on Computer Vision,* Part IV. Lecture Notes in Computer Science Volume 6314. Berlin: Springer.

Fang, H.; Gupta, S.; Landola, F. N.; Srivastava, R.; Deng, L.; Doll, P.; Gao, J.; He, X.; Mitchell, M. Platt, J. C.; Zitnick, C. L.; and Zweig, G. 2015. From Captions to Visual Concepts and Back. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: Institute for Electrical and Electronics Engineers. dx.doi.org/10.1109/CVPR.2015.7298754

Gao, H.; Mao, J.; Zhou, J.; Huang, Z.; Wang, L.; and Xu, W. 2015. Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question Answering. Unpublished paper deposited in The Computing Research Repository (CoRR) 1505.05612. Association for Computing Machinery.

Geman, D.; Geman, S.; Hallonquist, N.; and Younes, L. 2015. A Visual Turing Test for Computer Vision Systems. *Proceedings of the National Academy of Sciences* 112(12): 3618–3623. dx.doi.org/10.1073/pnas.1422953112

Hodosh, M.; Young, P.; Hockenmaier, J. 2013. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *JAIR* 47: 853–899.

Karpathy, A., and Fei-Fei, L. 2015. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: Institute for Electrical and Electronics Engineers. dx.doi.org/10.1109/CVPR.2015.7298932

Kiros, R.; Salakhutdinov, R.; and Zemel, R. 2015. Unifying Visual-Semantic Embeddings with Multimodal Neural Language. Unpublished paper deposited in The Computing Research Repository (CoRR) 1411.2539. Association for Computing Machinery.

Kulkarni, F.; Premraj, V.; Dhar, S.; Li, S.; Choi, Y.; Berg, A. C.; and Berg, T. L. 2011. Baby Talk: Understanding and Generating Simple Image Descriptions. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: Institute for Electrical and Electronics Engineers. dx.doi.org/10.1109/cvpr.2011.5995466

Lin, T. Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision–ECCV 2014: Proceedings of the 13th European Conference*, Part V. Lecture Notes in Computer Science Volume 8693. Berlin: Springer.

Malinowski, M., and Fritz, M. 2014. A Multi-World Approach to Question Answering about Real-World Scenes Based on Uncertain Input. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014,* 1682–1690. La Jolla, CA: Neural Information Processing Systems Foundation.

Mao, J.; Xu, W.; Yang, Y.; Wang, J.; Huan, Z.; and Yuille, A. L. 2015. Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN). Unpublished paper deposited in arXiv. arXiv preprint arXiv:1412.6632. Ithaca, NY: Cornell University.

Markoff, J. 2014. Researchers Announce Advance in Image-Recognition Software. *New York Times,* Science Section (November 17).

Mitchell, M.; Han, X.; Dodge, J.; Mensch, A.; Goyal, A.; Berg, A.; Yamaguchi, K.; Berg, T.; Stratos, K.; Daumé, H. 2012. Midge: Generating Image Descriptions from Computer Vision Detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computa-*

tional Linguistics. Stroudsburg, PA: Association for Computational Linguistics.

Richardson, M.; Burges, C.; Renshaw, E. 2013. MCTest: A Challenge Dataset for the Machine Comprehension of Text. In *EMNLP 2013: Proceedings of the Empirical Methods in Natural Language Processing Conference*. Stroudsburg, PA: Association for Computational Linguistics.

Tu, K.; Meng, M.; Lee, M. W.; Choe, T. E.; and Zhu, S. C. 2014. Joint Video and Text Parsing for Understanding Events and Answering Queries. *IEEE MultiMedia* 21(2): 42–70. dx.doi.org/10.1109/MMUL.2014.29

Vedantam, R.; Lin, X.; and Batra, T.; Zitnick, C. L.; and Parikh, D. 2015. Learning Common Sense through Visual Abstraction. In *Proceedings of the IEEE International Conference on Computer Vision,* ICCV 2015. Piscataway, NJ: Institute for Electrical and Electronics Engineers.

Vedantam, R.; Zitnick, C. L.; and Parikh, D. 2015. CIDEr: Consensus-Based Image Description Evaluation. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: Institute for Electrical and Electronics Engineers. dx.doi.org/10.1109/CVPR.2015.7299087

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and Tell: A Neural Image Caption Generator. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: Institute for Electrical and Electronics Engineers. dx.doi.org/10.1109/CVPR.2015.7298935

Weston, J.; Bordes, A.; Chopra, S.; and Mikolov, T. 2015. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. Unpublished paper deposited in arXiv. arXiv preprint arXiv:1502.05698. Ithaca, NY: Cornell University.

Zitnick, C. L., and Parikh, D. 2013. Bringing Semantics into Focus Using Visual Abstraction. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: Institute for Electrical and Electronics Engineers. dx.doi.org/10.1109/CVPR.2013.387

Zitnick, C. L.; Parikh, D.; and Vanderwende, L. 2013. Zero-Shot Learning via Visual Abstraction. In *Computer Vision-ECCV 2014: Proceedings of the 13th European Conference,* Part IV. Lecture Notes in Computer Science Volume 8692. Berlin: Springer.

Zitnick, C. L.; Vedantam, R; and Parikh, D. 2015. Adopting Abstract Images for Semantic Scene Understanding. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Issue 99.

C. Lawrence Zitnick is interested in a broad range of topics related to visual recognition, language, and commonsense reasoning. He developed the PhotoDNA technology used by Microsoft, Facebook, Google, and various law enforcement agencies to combat illegal imagery on the web. He received the Ph.D. degree in robotics from Carnegie Mellon University in 2003. In 1996, he coinvented one of the first commercial portable depth cameras. Zitnick was a principal researcher in the Interactive Visual Media group at Microsoft Research, and an affiliate associate professor at the University of Washington at the time of the writing of this article. He is now a research mananager at Facebook AI Research.

Aishwarya Agrawal is a graduate student in the Bradley Department of Electrical and Computer Engineering at Virginia Polytechnic Institute and State University. Her research interests lie at the intersection of machine learning, computer vision, and natural language processing.

Stanislaw Antol is a Ph.D. student in the Computer Vision Lab at Virginia Polytechnic Institute and State University. His research area is computer vision — in particular, finding new ways for humans to communicate with vision algorithms.

Margaret Mitchell is a researcher in Microsoft's NLP Group. She works on grounded language generation, focusing on how to help computers communicate based on what they can process. She received her MA in computational linguistics from the University of Washington, and her Ph.D. from the University of Aberdeen.

Dhruv Batra is an assistant professor at the Bradley Department of Electrical and Computer Engineering at Virginia Polytechnic Institute and State University, where he leads the VT Machine Learning and Perception group. He is a member of the Virginia Center for Autonomous Systems (VaCAS) and the VT Discovery Analytic Center (DAC). He received his M.S. and Ph.D. degrees from Carnegie Mellon University in 2007 and 2010, respectively. His research interests lie at the intersection of machine learning, computer vision, and AI.

Devi Parikh is an assistant professor in the Bradley Department of Electrical and Computer Engineering at Virginia Polytechnic Institute and State University and an Allen Distinguished Investigator of Artificial Intelligence. She leads the Computer Vision Lab at VT, and is also a member of the Virginia Center for Autonomous Systems (VaCAS) and the VT Discovery Analytics Center (DAC). She received her M.S. and Ph.D. degrees from the Electrical and Computer Engineering Department at Carnegie Mellon University in 2007 and 2009, respectively. She received her B.S. in electrical and computer engineering from Rowan University in 2005. Her research interests include computer vision, pattern recognition, and AI in general, and visual recognition problems in particular.